The Use of Survey Evidence in Intellectual Property Disputes

Professor Janet Hoek
Professor Philip Gendall

February, 2007

Massey University

The Law Foundation NEWZEALAND

The Use of Survey Evidence in Intellectual Property Disputes

ACKNOWLEDGMENTS

The research outlined in this report was generously funded by the New Zealand Law Foundation. We are grateful to Lynda Hagen and Joane Ciesiolka-Lord for their patience and support throughout the research process.

TABLE OF CONTENTS

1. Evolution and Status of Survey Evidence 1
2. Acceptability of Consumer Survey Evidence 6
2.1 Validity 6
2.2 Causality 7
3. Criteria Relating to Consumer Survey Evidence 9
3.1 Standards of Consumer Survey Evidence 9
3.2 Expert Evidence Relating to Survey Acceptability 11
4. Survey Error Sources 13
4.1 Coverage Error 14
4.1.1 Screening Questions 15
4.1.2 Coverage Error in Context 16
4.1.3 Geographic Coverage 21
4.1.4 Testing Coverage Error 22
4.1.5 Summary 25
4.2 Sampling Error 26
4.2.1 The Courts' Assessment of Sample Size 27
4.2.2 Assessing the Sampling Procedure Employed 30
4.3 Non-Response Error 33
4.4 Measurement Error 35
4.4.1 Question Wording 35
5. Establishing Secondary Meaning 63
5.1 Genericism and Distinctive Marks 63
5.2 Approaches to Estimating Secondary Meaning 65
5.3 Estimating Confusion 72
6. Depth Interviews with Intellectual Property and Survey Research Experts 93
6.1 Methodology 93
6.2 Research Findings: Academic Expert Witnesses 94
6.2.1 Characteristics of Legal Surveys 94
6.2.2 Coverage Error 95
6.2.3 Sampling Error 97
6.2.4 Non-Response Error 98
6.2.5 Measurement Error –Questionnaire Design 99
6.2.6 Measurement Error – Interviewer Quality Assurance 101
6.2.7 Role of Experts 103
6.3 Research Findings: Market Researchers 105
6.3.1 Coverage Error 105
6.3.2 Sampling Error 106
6.3.3 Non-Response Error 107
6.3.4 Measurement Error –Questionnaire Design 108
6.3.5 Measurement Error – Interviewer Quality Assurance 109
6.4 Research Findings: Lawyers 111
6.4.1 Role of Survey Research 111
6.4.2 Coverage Error 114
6.4.3 Sampling Error 115
6.4.4 Non-Response Error 117
6.4.5 Measurement Error –Questionnaire Design 118
6.4.6 Measurement Error – Interviewer Quality Assurance 120
6.4.7 Levels of Confusion and Association 121
6.4.8 Role of Experts 122
6.5 Conclusions 124
7. Conclusions 126
References 127
APPENDIX 1: Calculation of Error Margins 132
APPENDIX 3: Criteria to Guide Survey Development and Data Collection 136
APPENDIX 4: Interview Protocol 142

1. Evolution and Status of Survey Evidence

1. Traders have responded to more aggressive marketing environments by registering attributes they believe confer a competitive advantage on their brands. They have also monitored competitors' activities closely, and have taken action to restrain behaviour they believe may mislead consumers and ultimately disadvantage their brands. In particular, there has been a growing number of disputes over brand names, colours, logos, shapes, and other aspects of trade dress.

2. These disputes arise because of traders' concerns that consumers may confuse brands or mistake the source of these. Senders and Green (1998) noted that if human perception were perfect, there would be no opportunity to create misleading impressions, or to pass off one brand as another, since consumers would immediately detect unclear and misleading communications. Perception is arguably low in many purchase situations since routine purchases involve low levels of deliberation and attention, thereby creating a non-negligible potential for consumers to confuse brands that feature similar attributes.

3. When arguing that a trademark has been breached or that a rival has engaged in allegedly deceptive behaviour, companies have often submitted consumer survey evidence to support their claims. The acceptance of survey evidence marks a change in the courts' previous reliance on a judge's or hearing officer's interpretation of the case and its likely effect on relevant consumers. The introduction of survey evidence recognised that decisions may be informed by consideration of a wider range of consumer perspectives (Barksdale, 1959; Brandt and Preston, 1977). Keller (1992) summarised this view when he noted that surveys can constitute a powerful weapon that counsel may use to demonstrate a particular state of mind or situation. Clearly, if survey evidence indicates confusion exists among consumers; for example, if it reveals a higher than expected propensity to mistake one brand for another, then surveys could form a valuable part of cases alleging (or refuting) passing-off or deceptive conduct claims.

4. Initially, surveys were used to identify individuals who could subsequently be asked to appear as witnesses in court. However, this approach has obvious limitations since, for every witness willing to support the plaintiff's case, there is likely to be another who will testify in support of the respondent. The net effect of inviting compliant witnesses is an intellectual stalemate where neither side may succeed in establishing its case. More fundamentally, the representativeness of those witnesses prepared to provide evidence is questionable, since it would be unwise for counsel to call witnesses who they could not be confident would support their position. Thus, even if one side does produce more credible witnesses, judges may, for good reason, be reluctant to generalise their views to a wider population (Caughey, 1956).

5. The robustness of evidence provided by respondents recruited through a survey could be tested during cross-examination, which would probe the qualities of those providing opinions. Cross-examination could thus assess whether a concern Foster J noted, the issue of whether affidavits prepared by members of the public had been sworn by "*morons in a hurry*", was evident. However, testing the evidence provided by each individual consumer could be time-consuming, particularly if a large number of affidavits had been submitted (*Morning Star Cop-op Soc Ltd v Express Newspapers Ltd* [1979] 5 FSR 113 at 117).

6. To avoid the problem that evidence from selected consumers was not necessarily representative of a wider population, the results from consumer surveys were adduced. Survey evidence comprised responses elicited from a broader cross-section of relevant consumers and was not limited to those respondents likely to lend support to a particular position. However, consumer survey evidence was initially viewed as hearsay because the individuals interviewed were typically not available for cross-examination. As a result, the views and responses attributed to them could not be tested and were considered unreliable.

7. Furthermore, the opposing side sometimes had no direct knowledge of the interviewing methods used, the questions put to respondents, or the way in which responses to these questions were recorded. Without this information, counsel is unable to test the robustness and validity of the resulting estimates and the admissibility of evidence that cannot be scrutinised in this way has thus been challenged.

8. However, *Customglass Boats Ltd v Salthouse Bros Ltd* [1976] 1 NZLR 36, which examined the reputation of the name "Cavalier", involved a consumer survey that Mahon J accepted as valid evidence. He rejected the objection that the survey results were mere hearsay and observed that surveys could be of assistance in "*proving an external fact, namely that a designated opinion is held by the public or a class of the public*", which he stated was not a matter of hearsay at all. More generally, His Honour ruled that market surveys could prove "*a public state of mind on a specific question*", which he noted was an exception to the hearsay rule (para 1).

9. However, Mahon J drew an important distinction between the existence of opinions and the status of those opinions. He noted that while surveys establish that respondents held and offered the opinions recorded in the survey, surveys did not demonstrate the truth of those opinions ([1976] 1 NZLR 36 at 40; see also Barksdale, 1959, who discusses the US courts' stance).

10. Mahon J's comments confirmed that properly conducted surveys elicit views from a sample that is a microcosm of the wider population affected, or potentially affected, or at risk of being affected, by the behaviour of a trader. Surveys identify the proportion of a population affected by allegedly confusing behaviour and may also ascertain the likely effects of any deception. Fairly designed and administered surveys, and accurately recorded responses to these, can therefore offer important insights into how consumers interpret and respond to traders' behaviour (Keller, 1999).

11. Since *Customglass Boats Ltd v Salthouse Bros Ltd [*1976], consumer survey evidence has been adduced in several cases. While survey respondents may still appear as witnesses, their cross-examination is not normally designed to test the views they hold. Instead, cross-examination reviews whether the survey was properly conducted, the data appropriately interpreted and reported, and the conclusions advanced soundly based and fairly determined.

12. Yet, despite the growing acceptance of surveys, Deeth (2001) described US courts' response to survey evidence as paradoxical; while he noted that these courts were anxious to have consumer survey evidence (and would sometimes look unfavourably on its absence from submissions), he also noted that American judges were quick to reject surveys if they appeared flawed in any way. Although New Zealand judges have not lamented the absence of survey evidence in intellectual property disputes, they have occasionally commented on this point (see IPONZ case T36/2002; TM 224139 and *Carter Holt Harvey Ltd v Cottonsoft Ltd* [2004] 8 NZBLC 101, 588, paras 18 and 45).

13. McCarthy (1999) had earlier recognised the supporting role of surveys when he commented that researchers would be forgiven for concluding that judges accepted survey evidence when it reinforced the opinion they had already formed, and rejected it when it conflicted with their views. It is clear from recent New Zealand decisions that, while witnesses and surveys can assist the court, ultimately judges assess the dispute holistically; as Gendall J noted: *"the authorities emphasise it is for the Court to make up its own mind"* (*Austin, Nichols & Co Inc v Stichting Lodestar*, CIV 2004- 485-1281, para 19).

14. Moreover, survey evidence may not necessarily be considered to offer additional insights into a dispute. In *Noel Leeming Television Ltd v Noel's Appliance Centre Ltd* (1985) 1 TCLR 290, Holland J noted: *"I do not regard evidence from the survey as being significant either way in relation to confusion. In the end the issue is one for the Judge."* Although His Honour granted an injunction restraining use of the name "Noel's Appliance Centre", he declined to award costs of the survey against the defendant and accepted there was some credence in the argument that the survey was a *"sledgehammer* [used] *to crack a*

nut". The suggestion that survey evidence may be unnecessary was also advanced by Jacob J who noted: *"The Judge must consider the evidence adduced and use his own common sense and his own opinion as to the likelihood of deception. It is an overall jury assessment involving a combination of all these factors… ultimately the question is for the Court, not for the witnesses." Neutrogena Corporation v Golden Limited* [1996] RPC 432, p. 482.

15. Yet, despite discussion over the acceptability and role played by survey evidence, Deeth (2001) suggested surveys have three benefits that should prompt the courts to pay more attention to them: they are cost-effective, accurate and bring a realistic perspective to cases. However, while Deeth (2001) argued that consumer surveys were efficient and could save both time and money, since they avoided the need to present large numbers of consumers in person, the costs of designing a questionnaire, undertaking a rigorous survey, and preparing a careful and thoughtful interpretation of the survey findings are not inconsiderable. As a result, the perceived cost and time savings may be less than Deeth (2001) suggests, particularly if survey evidence is vulnerable to criticism from opposing counsel. Nevertheless, the alternative, which is to procure and cross-examine a range of witnesses, is clearly less efficient from the court's perspective.

16. Other benefits that Deeth (2001) attributed to survey evidence (its accuracy when compared to a parade of witnesses and its reality, or ability to bring, as he described, "*the real world into the courtroom*") have also been challenged. Thus, although Deeth (2001) suggested survey evidence could give a "*truer picture*" than an individual judge's assessment, his claim assumes judges accept that consumer surveys are sufficiently robust to offer insights that go beyond those they can offer from their own personal experience.

17. To support the role of survey evidence in providing a broad overview of consumers' opinions and behaviours, Deeth (2001) quoted from *Sun Life Insurance v Sun Life Juice* (1988) where the judge noted that lack of *"regard to evidence of what others may think or have said would to my mind be nothing more than an exercise in judicial fantasy".* While the courts have often accepted consumer survey evidence, it is clear that not all members of the judiciary have thought it more robust than "judicial fantasy". For example, Eko (1998) reported a similar view when he quoted a judge who claimed: "*statistics are elusive things at best, and it is a truism that almost anything can be proved by them*" (pp. 592-593).

18. Thus, despite the advantages Deeth (2001) argues consumer survey evidence provides, judges continue to view surveys with some scepticism and the weight attached to this evidence has varied considerably. Whitford J, in *Imperial Group plc v Philip Morris Ltd* [1984] RPC 293 at 302-303 noted: *"However satisfactory market research surveys may be*

in assisting commercial organisations as to how they can best conduct their business, they are by and large, as experience in other cases has indicated, an unsatisfactory way of trying to establish questions of fact which are likely to be maters of dispute." Jeffries J expressed similar views in *Comité Interprofessionel du Vin de Champagne v Wineworths Group Ltd (1991)* 2 NZLR 432, where he observed that *"market research methods are almost as arguable as some of the conclusions" 44*5 *(1991)* 2 NZLR 432.

19. Yet, notwithstanding judicial ambivalence about the merits of survey evidence, consumer surveys have been conducted to assist several types of intellectual property proceedings. First, they have been used in trademark registration cases, where applicants attempt to demonstrate that a particular mark has become distinctive and acquired secondary meaning. Second, they have been used by opponents to trademark applications; in these cases, the surveys have been used to demonstrate that a mark is in common use or is generic, and thus that its registration as a trademark would create a likelihood of confusion among the relevant public. Third, surveys have been used to demonstrate a likelihood of confusion following advertising, or the use of aspects of trade dress that, in aggregate, create an allegedly confusing impression. Finally, surveys have been used in passing off cases, where claimants have sought to demonstrate that advertising claims or a brand's appearance have led to confusion that has damaged the goodwill the claimant has in a brand. Survey evidence may thus provide insights into a range of cases; however, its acceptance and the weight eventually attached to consumer surveys will depend critically on their design, implementation and interpretation (Eko, 1998). The following section examines some of the characteristics survey evidence must possess before it is accepted.

2. Acceptability of Consumer Survey Evidence

2.1 Validity

20. In evaluating the role survey evidence may play in court proceedings, researchers have raised both philosophical and methodological questions about its validity. Those raising more fundamental issues have questioned the acceptability of survey evidence and the extent to which it can inform decisions, while those concerned with methodological issues accept the admissibility of survey evidence but debate the probative weight it should have. Eko (1998) raised several questions relating to the validity of consumer survey evidence. He began by arguing that this evidence is not objective, since by measuring a variable, researchers intervene in an environment and thus alter the variable they wish to estimate. This criticism recognises that surveys are intrusive and that, by drawing attention to a particular behaviour, or the factors that may influence that behaviour, researchers may increase the salience of those factors, thereby influencing the behaviour they wish to investigate.

21. However, while the intrusiveness of surveys is undeniable, this need not necessarily affect the variable of interest. Where the behaviours measured are subject to social approval, survey evidence may be biased if efforts have not been made to address social desirability error. However, measurement of routine behaviours that are frequently performed seems less likely to be altered as a consequence of being measured. That is, where behaviours are regular, where the level of thought expended on the decision is not extensive, and where consumers are accustomed to the cues prompting these, the introduction of those cues as part of a survey or experiment would seem unlikely to exert undue influence on consumers' responses.

22. Notwithstanding the robustness of habitual behaviours to measurement, concerns over the ability of surveys to elicit valid responses have also been raised by judges themselves. For example, US Chief Judge Posner of the 7th Circuit raised interesting questions about the ability of surveys to replicate real world conditions. He suggested that people tended to be more cautious when "*laying out their money*" than they were when answering survey questions, which they had no incentive to answer accurately or even honestly.

23. Establishing the validity of survey evidence is critical and it is clear that, in some cases, surveys have been viewed as unscientific and subsequently described as invalid. For example, in *CIBA-GEIGY v Douglas Pharmaceuticals Ltd* (1986) 2 TCLR 346, a case involving a request for extension to the term of a patent, survey evidence relating to specialist rheumatologists' views on a drug was adduced. However, an expert charged with evaluating this evidence found that the covering letter outlined the desired responses, less than half (14/34) the questionnaires were returned, and respondents had

appended notes to the survey indicating concerns over the validity of the evidence collected (1986) 2 TCLR 359. Assistant Commissioner Burton declared he had *"no hesitation in reaching the conclusion that Dr Stephen's survey is of no assistance whatsoever"* (1986) 2 TCLR 359.

24. While naively-worded and poorly conducted surveys may not elicit meaningful responses, and while rogue respondents may exist, surveys can be designed to create a context that simulates how respondents would perform the behaviour in question. In addition, individuals who agree to participate in a survey typically answer the questions put to them as best they can, otherwise they would have exercised their right to decline to co-operate. Overall, surveys can be designed to simulate the conditions that prevail when respondents "*lay out their money*"; this creation of settings that are analogous to real world conditions increase the probability that the data elicited have external validity.

2.2 Causality

25. Eko (1998) also argued that surveys do not establish causality; as a result, he argued researchers cannot prove that exposure to a particular claim, trademark, or advertisement led to or created confusion. This argument has merit, but overlooks a more fundamental question about what the courts require. Evidence of causality is notoriously difficult to obtain in any social science setting, thus researchers examine correlations, the strength of these, and the consistency with which they appear. As a result, although the cross-sectional surveys normally used in evidence cannot establish causality, since they measure behaviour or opinions at a particular point in time and do not identify variables that cause behaviour, they can identify factors that correlate with behaviour. To establish causality would require longitudinal studies that enable the roles of different factors on variables of interest to be estimated. In intellectual property cases, few parties to a dispute would have access to longitudinal data. Moreover, as forensic surveys are designed to explore a specific legal question, they are invariably cross-sectional and represent views at the particular date on which confusion allegedly occurred, or on which a trademark application was made.

26. As a result, few cases have drawn on or seen adduced longitudinal studies where respondents' behaviour has been measured over time and where imputations of causality become possible. However, although establishing causality could strengthen the case being advanced, it is not necessary to demonstrate that deception has occurred. Instead, researchers need to test whether the disputed attributes or claims are likely to mislead or deceive consumers; this typically involves assessing respondents' interpretations of the allegedly deceptive attributes and the beliefs they come to hold about these. Cross-sectional studies are an appropriate methodology for establishing consumers' cognitive responses to brand attributes, thus, despite Eko's arguments to the contrary, the need to

present longitudinal survey evidence that demonstrates causal associations would not seem necessary.

27. Although Eko's (1998) general criticisms of survey evidence can be refuted, his comments highlight the wide range of criticisms that affect the admissibility of consumer survey evidence adduced in New Zealand and the weight attached to this. As Gastwirth (2003) also noted, it is nearly always possible to point to flaws in consumer survey evidence because surveys, no matter how carefully designed, implemented or interpreted, contain potential error since the estimates are based on a sample rather than a population. As a result, all survey estimates are affected by sampling error, the size of which depends on the sample size and the selection procedure used to recruit respondents. We consider these methodological issues in more detail in subsequent sections, but turn now to the question of how counsel and those hearing a case may assess criticisms of survey evidence and how these may affect the conclusions advanced. To assist with this process, it is helpful to have a set of criteria to which consumer survey evidence should conform.

3. Criteria Relating to Consumer Survey Evidence

3.1 Standards of Consumer Survey Evidence

28. Because the quality of consumer survey evidence adduced in the United States was initially variable, a judicial study group developed criteria they recommended surveys should meet if they were to be considered acceptable. These were subsequently adopted and reinforced in Federal Rule of Evidence Rule 703, which was issued in 1975, and they have provided detailed guidance to counsel wishing to adduce survey evidence. Adherence to these criteria led Sarel and Marmorstein (2002) to observe that, in the intervening three decades, *"the courts have moved from resistance to mild acceptance and encouragement of the presentation of survey evidence. Courts now routinely criticise litigants for failure to present survey evidence when appropriate"* (p. 12). They go on to note that survey evidence is now expected to be used to demonstrate a likelihood of confusion and is also widely used in establishing secondary meaning where trademark registration is sought.

29. A US case *Toys R Us v Kanasarie Kiddie Shop* 559 F. Suppl. 1189, 1205 EDNY 1983, was among the first to outline criteria that the courts explicitly considered when reviewing survey evidence. These included the need to:
    - provide a proper definition of the survey universe,
    - draw a representative sample from this universe,
    - use interviewers who present questions in a clear, precise and non-leading manner,
    - withhold information from both interviewers and respondents about the purpose for which the survey was being conducted,
    - report the data accurately,
    - analyse the data according to accepted statistical principles, and
    - provide evidence that the survey design and implementation had been objectively developed.

30. In the United Kingdom, *Whitford J in Imperial Group plc v Philip Morris* [1984] RPC 293 set out nine principles that have guided the development and use of consumer survey evidence in New Zealand. These principles relate to the sample size and selection; the design and conduct of the survey, and the level of disclosure of the survey design and administration required, and are very similar to those Mahon J relied on in *Customhouse Boats Ltd v Salthouse Bros Ltd* (Customhouse Boats) [1976] 1 NZLR 36.

31. Referring to US cases, Morgan (1990) presented a set of standards that included both explicit or *"patently obvious"* criteria and those he described as implicit rules, which arose from earlier decisions (p.59). Again, these criteria are similar to the judicial

guidelines set out by Whitford J in the UK *Imperial Group* case. Similarly, the Federal Court of Australia has issued a Practice Note dealing with survey evidence (No. 11, 1994) and some analyses of Australasian cases have been undertaken (Skinnon and McDermott, 1998).

32. Guidelines issued by Canadian judges have contained very similar criteria. Mr. Justice Mackay noted several guidelines in *Joseph E. Seagram & Son Ltd.* v. *Seagram Real Estate Ltd.* (1990), 33 C.P.R. (3d) 454 (F.C.T.D.). In particular, he found that acceptance of a survey and the weight attached to it would depend on:
    - the time period in which the survey took place;
    - the questions asked;
    - where they were asked, and
    - the method of selecting participants.

33. In addition, His Honour noted that the researcher involved in the survey should provide an affidavit containing details of how the questionnaires were administered and completed. Further guidelines included the desirability of using natural rather than artificial environments; the need to conduct the survey at or around the time of the trademark application or opposition; and the need for a representative sample. His Honour also suggested that open ended questions were preferable to closed questions because the former did not prompt respondents' answers in any way (Grenier, undated).

34. New Zealand courts have also recognised the need to establish standards that consumer survey evidence should meet. Thus, as well as clarifying the acceptability of consumer survey evidence, Mahon J set out specific criteria that surveys should meet before the courts could accept them. His Honour set out three criteria that determined the weight accorded survey evidence:
    - The use of non-leading questions (*"questions formulated in such a way as to preclude a weighted or conditioned response")* (para 2);
    - Evidence of accurate transcription (*"clear proof that the answers were faithfully and accurately recorded")* (para 2);
    - Collection of information from an appropriate sample (*"evidence that the answers were drawn from a true cross-section of that class of the public or trade whose impression or opinion is relevant to the matter in issue")* (para 2).

35. Overall, His Honour concluded that survey evidence that met these criteria was preferable to the "*interminable parade of witnesses*" otherwise called to demonstrate wider public opinion on a particular issue *Customhouse Boats Ltd v Salthouse Bros Ltd* (Customhouse Boats) [1976] 1 NZLR 36.

36. In *Noel Leeming Television Ltd v Noel's Appliance Centre Ltd* (1985) 1 TCLR 283, Holland J summarised Whitford J's observations in *Imperial Group plc v Philip Morris Ltd* [1984] RPC 293. His Honour noted that surveys must select respondents to ensure they are representative of the population of interest; that the sample size must be adequate for the analyses performed, and that the survey itself must be fairly conducted. In order to establish that the survey has been fairly and properly conducted, those adducing it must provide what Whitford J described as: *"the fullest possible disclosure of exactly how many surveys they have carried out, exactly how those surveys were conducted and the totality of the number of persons involved, because otherwise it is impossible to draw any reliable inference that answers given by one or two or three people in one survey might conceivably be said to indicate that similar answers would be given of a survey covering the entire smoking population."* (1985) 1 TCLR 287.

37. To demonstrate that survey evidence has met these standards, legal counsel has engaged experts to review and evaluate the evidence they wish to adduce. Where experts have endorsed a survey design and confirmed that the survey was fairly administered and the data appropriately interpreted, the courts have tended to accept the survey, unless compelling expert evidence submitted by opposing counsel leads them to question the survey design or conclusions. The following section examines the use of expert witnesses and the evolution of their role in attesting to the validity and reliability of consumer survey evidence.

3.2 Expert Evidence Relating to Survey Acceptability

38. The growing use of experts to assist with survey design and to assess the conclusions drawn from the survey data has also led to the development of rules that judges apply to experts themselves. Deeth cites from *Cordon Bleu v Bradley* 1979, where the judge noted *"It is well understood that the expert must be qualified and that the survey must be conducted according to the rules of the art, that the compilations must be verified and that the results must also be interpreted according to the experience and the basic rules of the science and the art. The rest is a question of probative value."* This reinforces the view that, while experts can support the acceptability of a survey, the question of its weight remains a matter that judges and hearing officers determine.

39. Experts usually hold senior academic positions or management positions in market research companies, and rules that relate to their conduct have recently been established both in New Zealand and internationally. *Daubert v Merrell Dow Pharmaceuticals, Inc.* 509 US 579 (1993), saw a revision of the test applied to the admissibility of expert evidence. Hudis (2000) summarised the US courts' rules for expert evidence, which include:

- Whether the theory or technique the expert seeks to introduce has been or is able to be empirically tested;
- Whether the theory or technique has been subjected to peer review;
- Whether the theory or technique has a known rate of error;
- Whether the theory or technique is generally accepted in the scientific community.

40. Although the Daubert case applied only to scientific evidence, subsequent cases broadened this to include other forms of expert testimony (*Kumho Tire Co. v Carmichael*, 119 S.Ct. 1167 (1999). Hudis (2000) noted that consumer surveys, whether used to test the likelihood of confusion between two brands, or to test the genericness of a mark, are also "*ripe candidates*" for the courts' gate keeping scrutiny.

41. However, Hudis (2000) also commented on the varied reception survey evidence that experts have designed, interpreted and presented has met. While one judge described it as offering important insights, another considered it expensive and open to manipulation. Interestingly, in Indianapolis *Colts Inv,. V Metropolitan Baltimore Football Club Limited Partnership*, 34 F.3d 410 (7th Circuit, 1995), the court suggested that the experts representing each side in the dispute should be asked to nominate a third, neutral expert, who could conduct the proposed studies. These comments recognise the value of evidence that is disinterested, and that has been conducted for the court, rather than for one of the parties to the dispute.

42. Lessem (2000) noted concerns over the disinterestedness of experts when he pointed out that although the courts have generally accepted survey evidence, some judges have described this as "*less than edifying*", largely because of the way experts interpret the survey evidence they have helped to design. He quoted Chief Judge Posner of the 7th Circuit: "*Many experts are willing for a generous (and sometimes for a modest) fee to bend their science in the direction from which their fee is coming. … The judicial constraints on tendentious expert testimony are inherently weak because judges… lack training or experience in the relevant fields of expert knowledge*" *Indianapolis Colts, Inc. v. Metropolitan Baltimore Football Club L.P.*, 34 F.3d 410 (7th Cir. 1994).

43. Ford (2005) noted that recent US decisions had led to a re-appraisal of experts' role and a review of the types of surveys considered acceptable. Criteria used to assess scientific evidence have also been applied to survey evidence. For example, Ford noted that judges pay particular attention to the qualifications of the expert commenting on or assisting with the design of survey evidence. He also argued that experts' approach to a survey must be demonstrably similar to other work they have conducted and challenges that an expert departed from her or his normal practice would reduce the weight attached to the survey.

44. In addition, Ford outlined the importance of using methodologies and explanations that intelligent lay audiences can understand and review. He argued that experts should clearly outline the methodologies they employed in a survey design, but should ensure that the explanation advanced is lucid and accessible. Furthermore, he suggested that the methodologies used should be well-established and have a demonstrated history of validity that can be drawn on to support their use.

45. Ford's comments highlight the increasing scrutiny applied to scientific evidence, including survey evidence. In particular, details of the questions asked, the population surveyed, the sample selected, the interviewers' behaviour, and the interpretation and presentation of the results, have all been called into question and affected the weight given to a survey.

46. Experts have generally recognised that evidence adduced in court will be subject to more detailed and critical scrutiny than other research (Crespi 1987; Morgan 1990; Maronick 1991). While marketers may be highly experienced in commissioning surveys to measure awareness following an advertising campaign, or preference for a range of possible brand attributes, these latter surveys rarely if ever attract the level of scrutiny directed at forensic surveys. Overall, this lack of experience in conducting surveys for presentation in Court, together with a failure to draw on relevant international information and reluctance by litigants to fund high quality work, has resulted in surveys that provide little guidance and carry little weight. Halford-Harrison and Perkins (2004) summarised this state of affairs when they described surveys as "*Expensive tool[s] which, at best, [are] only persuasive and, at worst, can support the defendant's case*" (p. 2).

47. To ensure survey evidence withstands critical scrutiny, researchers attempt to design surveys unaffected by errors that will render the survey findings inadmissible, or of little assistance. This task is complex, since surveys are almost inevitably affected by error that may arise from four main sources. The following section examines these four types of error and considers how they might be addressed.

4. Survey Error Sources

48. Although trademark and fair trading law may differ from one jurisdiction to another, principles of sound survey design have a universal application. Dillman (2000) defined four types of error that may affect the validity and reliability of survey estimates: coverage error, sampling error, non-response error, and measurement error. In the remainder of this section, we review the attention paid by the courts to these errors,

suggest how flaws that fatally undermined the survey evidence could have been avoided, and outline principles that might guide researchers charged with designing forensic surveys.

4.1 Coverage Error

49. Coverage error is the error that occurs when there is a discrepancy between the sample and the wider population whose behaviour is of interest. Dillman defined coverage error as a measure of how adequately the survey sample represents the population of interest; it assesses whether "*every unit in the survey population* [has] *a known, non-zero chance of being included in the sample"* (Dillman 2000, p. 196). Defining the population of interest clearly and appropriately may reduce coverage error, as may the use of a sampling frame that ensures each member of that population has a known (non-zero) chance of inclusion in the sample. According to Morgan (1990), coverage error also refers to the need to ensure that the sample includes only relevant respondents, whom he defined as those able to address the legal question of interest.

50. When considering trademark disputes, Folsom and Teply (1998) noted that surveys must be selected from the "*relevant consuming universe*". That is, a sample should be "*all those consumers who may ultimately rely on the trademarked word in making purchasing decisions. It should include both past and prospective buyers, and each distinct subgroup of purchasers should be considered and perhaps surveyed separately"* (p. 12). The FTC Manual for Complex Litigation also notes the importance of ensuring the relevant population is properly chosen and defined.

51. The FTC Manual stresses the importance of ensuring that the sample selected is representative of the overall population of interest. Many survey researchers routinely weight their samples to ensure these match known population parameters, most commonly, the age and gender profile of the population (information that can easily be gleaned from Census data). However, this weighting is not necessarily sensible in the case of forensic research as demographic variables may be only weakly correlated with the behaviour of interest. For this reason, selecting a sample that fairly and accurately represents the wider population may require more sophisticated sampling procedures, such as stratification, to ensure that the sample includes sufficient respondents from sub-groups, thus enabling separate analysis of these. However, as Deeth (2001) noted, drawing a sample that enables assessment of the views of a range of sub-groups requires a more complex survey design and a larger sample, both of which can increase the overall cost of the survey.

52. To clarify how these suggestions might be implemented, Dupont (2001) noted that the first question to be determined in a survey is who is likely to be confused; he suggested

this group could include actual and potential consumers of all parties to the dispute. In practical terms, this may mean researchers need to select a broad sample that will enable them to analyse the responses of different sub-groups that constitute the overall population. Aggregation of the overall responses may mask variations that exist between sub-groups. For example, vulnerable groups that are more prone to accepting misleading claims may exist within a broader population; if the data analysis does not include breakdowns by respondents' education or other demographic variables, the effect of the allegedly deceptive claims on these groups may not be revealed. Analysis of sub-groups is thus necessary to obtain insights into the level and type of confusion that occurs within the sample (and the population this represents).

53. Consumers' behaviour may also affect their perception of claims. Thus responses to allegedly deceptive stimuli may vary depending on whether consumers see the brand as the dominant brand in their repertoire (their preferred or main brand), or whether the brand is a secondary brand that constitutes a smaller proportion of their share of category requirements. Drawing a sample that includes appropriate proportions of consumers from these groups is thus crucial to ensuring the survey's usefulness.

54. Deeth (2001) recognised the existence of different consumer groups when he suggested researchers need to go beyond examining the views of current product or service users and argued that the relevant universe contains not only actual users of both the defendant's and plaintiff's products, but also potential users and possibly even non-users. He stressed the need to consider carefully how *potential* users can be included in the survey sample, since failure to identify and survey this group may lead to rejection of the evidence. One approach to this issue involves the use of screening questions, which we discuss in more detail below.

4.1.1 Screening Questions

55. Screening questions can be used to exclude those with no interest in or experience of the product category. However, even respondents with a low probability of ever using the product category could be considered potential future users and should arguably be included on the grounds that some may subsequently develop an interest in the product category. Careful thought should therefore be given to decisions that would exclude them from the processing sample. Thus, while screening can exclude people whose views are based on neither knowledge nor experience, the screening criteria must remain sufficiently loose to ensure those with a potential interest in the product category are not excluded. This requires a fine balance to avoid allegations that the sample has been deliberately biased in favour of a particular outcome, and highlights the importance of undertaking a variety of analyses that examine responses from the aggregate sample as well as the sub-groups that constitute it.

56. If used too overtly, very specific screening questions may alert respondents to the overall purpose of the survey and thus may sensitise them to the issues under consideration; this may give rise to allegations of imbalance when the survey is scrutinised by opposing counsel. For this reason, screening questions are sometimes left to the end of the survey, even though this means that respondents whose opinions are not relevant are still interviewed. Those funding the research may view this practice as wasteful, since respondents who fail to meet the screening criterion may not be included in the processing sample. Irrespective of the method of screening adopted, researchers need to undertake careful analyses of sub-groups, which can be defined according to their use of the brand in dispute, their use of rival brands, and their overall experience of the product category.

57. Where a range of analyses is undertaken, researchers have the opportunity to demonstrate that the analysis was rigorously conducted and applied to a range of sub-groups. However, where the opinions and behaviours of these sub-groups differ, the researchers may find themselves having to explain apparently discrepant results. While consistent findings across a range of sub-groups can support the findings advanced, since these would seem to apply irrespective of the variation examined, it is also possible that analyses of sub-groups may reveal differences that are ultimately unhelpful. The decision about which groups to sample and examine in more detail has proved problematic in recent cases, we examine issues arising from this decision in the following section.

4.1.2 Coverage Error in Context

Defining the Population of Interest

58. The first task researchers face in addressing coverage error is to ensure they have defined the population of interest correctly. In some cases, they may identify more than one specific group whose perceptions and behaviours are of interest. For example, views from both trades people and end-users may be relevant, particularly in trademark disputes where consumers may rely on information given to them by retailers. The need to consider the range of populations likely to be affected was evident in *Wineworths Group Ltd v Comité Interprofessionel du Vin de Champagne*(1992) 2 NZLR 327, where confusion over the use of the word "champagne" could occur among both final consumers and the retail trade.

59. Similar coverage issues arose in *Patience & Nicholson (NZ) Ltd v Cyclone Hardware Pty Ltd* (2001) 3 NZLR 490, which also involved surveys of different groups: retailers and end-users. However, while the survey adduced by P&N was careful to select from both groups, counsel for Cyclone criticised it, arguing that the retailers selected were largely P&N customers, thus biasing the estimates obtained. Instead, Cyclone argued that the

list should have been stratified by market share, to ensure appropriate representation of different retailers; counsel for Cyclone relied upon market share information to demonstrate how the sample was biased. However, Hansen J found the market share information provided by both sides confusing and determined that he did not have to assess the relative merits of the market share statistics. Instead, he considered that the issue facing him was whether the list used was adequate, and he concluded that because the list included a good representation of the various market segments, it was acceptable (para 88).

60. Cyclone raised the same concerns about the end-user sample used in the P&N survey, which they argued comprised consumers who purchased from a common range of drills that were mainly supplied by P&N. Thus, instead of obtaining a representative sample of end-users, the sample over-represented P&N users. However, Hansen J ruled that although this flaw reduced the value of the survey evidence, it did not render it inadmissible. He considered that the survey *"may be used provided that it is accurately represented and appropriately qualified"* (para 91).

61. Similar concerns were documented in a recent UK decision, *Black and Decker v Atlas Copco Aktiebolag* (2003). In this case, Black and Decker had sought to register the colours black and yellow as they applied to electrically powered hand tools. They argued that consistent use of these colours in their DeWalt range had led the colours to become distinctively associated with this brand, and they adduced survey evidence in support of their application. However, the survey used to gather evidence that could be used to support this application was criticised on several grounds.

62. In particular, the sample was criticised for being unrepresentative. The surveys were conducted at specialist tool outlets, even though the relevant population was defined more broadly as the general power-tool purchasing public. Consumers who frequented specialist power tool outlets were more likely to be members of the trade than members of the public. It was also not clear how the outlets at which the interviewing took place had been selected and the courts questioned whether the outlets used had been randomly selected to be representative of the wider range of such stores.

63. The UK Black and Decker survey was also criticised for the way in which respondents were selected. Some of those interviewed worked in the trade, including the store at which the interviews were conducted. Because of their atypical levels of exposure to the brand in question and the promotional materials relating to that brand, their ability to represent ordinary consumers was highly questionable. Use of screening questions that identified respondents' type of usage (trade or "do-it-yourself") could have resolved these problems, as could a directive that store staff were not to be interviewed as part of

the general consumer study. This survey highlighted the need to ensure the interview sites used were appropriate and did not favour one party and that the selection methods used to identify eligible respondents were fair and defensible.

64. These cases highlight the need to consider affected populations more generally, and to ensure that, once these are defined, the samples are drawn in a fair and defensible manner. Hansen J did not specify what appropriate qualifications should have been applied to the Patience & Nicholson survey evidence, but the use of a sampling frame that does not ensure all users of the product category have a non-zero chance of being selected to participate would appear likely to have a material bearing on the estimates obtained. Similarly, failure to exclude respondents who have knowledge and experiences that mean they are unlikely to represent those at risk of confusion will also detract from the rigour and value of the survey evidence.

Identifying Relevant Populations

65. Other cases have attracted criticism because the dispute has affected more than one group within a market, but members of one groups only have been interviewed because they have been assumed capable of representing the views of other potentially affected users. The inadequacy of this approach was made clear in Re *Estheal; Pierre Fabre SA v Estée Lauder Cosmetics* (1989) 3 TCLR 133. Here, evidence that took the form of surveys administered to counter staff and retailers was adduced by Estée Lauder to provide insights into the likely behaviour of end users of perfumes and cosmetics.

66. The survey questions examined respondents' awareness of the marks Estée Lauder and Estée, the length of time they had been familiar with these marks, their association of these marks with a particular source, and their knowledge of promotional activities undertaken to support products bearing these marks. In addition, respondents were asked about their awareness of Estheal and their likely response, should they encounter products bearing this mark; of the eight respondents, 6 were not aware of Estheal. When asked how they would respond if they encountered products bearing this mark, respondents were generally rather cynical; many suggested the mark was trying to benefit from the cachet associated with the Estée Lauder name.

67. A further question explored whether respondents thought they, their staff, or other people, could be confused should the three marks (Estée Lauder; Estée and Estheal) be available at the same retail outlet. Although respondents thought they and their staff would be unlikely to confuse the marks, since they received detailed training about the brands they retailed, they felt the public could be confused.

68. This evidence was criticised for several reasons, and counsel for Estheal argued that the evidence failed to meet standards required of survey research. For example, no information about the size of the group invited to participate in the survey or the protocols used to ensure the survey was fairly designed and conducted (criteria set out in earlier decisions) were provided. Perhaps more importantly, counsel for Estheal noted that the evidence sought the views of members of the trade, rather than members of the public, and he questioned whether retail assistants or counter managers could provide insights into whether consumers would or would not be confused by the introduction of an Estheal mark. Assistant Commissioner Martin agreed with the latter point and commented that the evidence was not so much a survey as it was a collection of views from individuals involved in retailing cosmetics and perfumes (see also the approach used in IPONZ case 25/1997, where pharmacists were asked to provide opinions on the way in which their customers referred to a hair care brand featuring the words "Gentle Care" in its name).

69. The same types of issues arose in IPONZ 625055, T51/2002), which involved an application by Robocup Federation to register the name Robocup; to support this application, a survey of video store owners was adduced. The survey evidence in this case was criticised because the data were collected not from end users (those at risk of confusing the name) but from store owners. Assistant Commissioner Walden found that wrong population had been surveyed and that the survey offered no insights into the views held by ordinary consumers, since they were not included in the sample.

70. These outcomes reinforce the need for researchers to consider the range of populations that may be affected by alleged deception and to ensure the views of each group are represented directly. Reliance upon trade members' view of how end consumers may behave seems likely to attract criticism that the evidence is speculative, particularly if it is unsupported by direct evidence from the group in question (see also *Unico Trading PTE Ltd v PT Indofood Sukses Makmur*, AP308/01, para 25). Depending on the size of the populations, ensuring the relevant groups' views are canvassed may require a range of different surveys, and these may involve different research methodologies.

Population Sub-Groups and Market Partitions

71. However, as well as considering how the populations are defined, researchers need also to consider whether specific groups within the population are more or less relevant to the issue being determined. Even where the correct general population has been identified, failure to recognise specific groups within this may also create coverage error and diminish the value of survey evidence. For example, *Anheuser Busch v Budejovicky Budvar National Corporation* [2001] 3 NZLR 666, which concerned a dispute over the name "Budweiser", involved a survey of individuals who had recently purchased or consumed packaged beer. However, market evidence suggested the beer market at that time comprised at least three partitions, and the relevant population was not beer drinkers in general, but consumers of premium priced imported packaged beers. The sample contained only a small proportion of respondents from this group, and thus arguably failed to provide insights into the likely effects of alleged confusion on the behaviour of actual purchasers from this partition.

72. In reviewing this criticism of the survey, Doogue J commented that coverage error was a technical rather than substantive issue; "*These sorts of criticisms … undermined rather than destroyed what little relevance the survey had*." 692 [2001] 3 NZLR 666. This is consistent with *Patience & Nicholson (NZ) Ltd v Cyclone Hardware Pty Ltd* (2001) 3 NZLR 490, where the flaws in the sample were not considered fatal flaws that would preclude consideration of the survey evidence.

73. Issues that concern how the population of interest has been defined and whether it comprises ordinary consumers or consumers who have more specialised interests, knowledge and behaviours, have arisen in many cases. For example, in *Automobile Club de L'Ouest, Aco v South Pacific Tyres New Zealand Limited* CIV 2005 485 248, Wild J referred to criticisms of the Club's surveys that alleged these were not gathered from the correct market. Here, the three surveys adduced had sampled ordinary consumers selected in a street intercept survey, whereas the correct sample, South Pacific Tyres argued, should have comprised purchasers of performance tyres (paras 50 and 115). Although this criticism was not fatal, it reduced the credibility of the survey evidence and thus the weight attached to it.

74. Similar concerns were raised in *Cookie Time Ltd v Griffins Foods Ltd* (2000: M1756/SW00), which involved a dispute over packaging used to present biscuits. Cookie Time presented a survey to support its contention that the packaging used by Griffins was deceptively similar to their own. However, counsel for Griffins alleged that the survey had not sampled the correct cross-section of the public, since those at risk of confusion purchased biscuits from direct sellers who called at their workplace. These individuals, it was suggested, were unlikely to have been available to participate in street or

supermarket intercept surveys conducted during the day. This argument highlights the fact that researchers must consider two related factors: first, that they have identified the appropriate population, and second, that they have ensured the time period over which the data are collected is such that a wide cross section of this population will be available to participate in the survey.

75. Questions about the relevant population and how this was defined were raised in *Blenhaven; International Cellars (Marlborough) Ltd v Montana Wines Ltd* (1989) 3 TCLR 115. While the population of interest could have been defined as wine drinkers, a group that would include people under the age of 18, it could also have been more strictly defined as wine purchasers, a group that could only include people aged at least 18 years or over. However, use of the former definition was not held by Assistant Commissioner Martin to be a serious failing for two reasons. First, he commented that wine drinkers aged under 18 would have been likely to know of and have sampled *Blenheimer,* the brand owned by the opponent. Second, he commented that the proportion of people in the sample whose age fell between 16 and 18 *"would not have been great"* (1989) 3 TCLR 124.

76. These comments suggest usage, rather than ability to purchase, was considered more important, perhaps because confusion may also occur in a usage context, and thus may affect even those unable to purchase the product directly. However, the second comment seems at odds with the first reason given as it implies that the inclusion of people unable to purchase the product may not have been appropriate, but is unlikely to have had a material effect on the estimates produced. If this interpretation is correct, it raises questions about when the inclusion of non-purchasers would become a flaw that undermined the credibility of the survey evidence.

4.1.3 Geographic Coverage

77. In *Cookie Time Ltd v Griffins Foods Ltd (2000: M1756/SW00)*, Griffins had also argued the Cookie Time survey was flawed as the date were collected only in Auckland and Wellington. No interviews were conducted in Christchurch, the city in which Cookie Time originated (and where Griffins presumably thought any confusion would be less evident). The question of the appropriate level of geographic coverage has arisen in a number of cases.

78. In *Levi Strauss & Co Ltd v Kimbyr Investments Ltd* [1994] 1 NZLR 332 counsel for Kimbyr argued that the survey was inadmissible, since the data had been collected from seven cities, which accounted for 55% of the total population. However, Williams J did not accept that all survey samples should be drawn from throughout the country and found that the survey was *"sufficiently representative"* [1994] 1 NZLR 332 at 364. This case

highlights the fact that researchers must both define the population correctly and ensure the sample is drawn from a sufficient number and variety of sites; failure to address these points will lead to allegations that the sample is not representative. Thus, while survey evidence need not be based on a nationwide sample, the data collection venues must provide access to a fair cross-section of the population of interest.

79. Because each case will involve a different population, it is difficult to provide prescriptive guidelines about the number of sites likely to be considered adequate. However, reference to the market structure, key retail outlets, and the distribution strategy employed by participants within that market will help determine where the interviewing should be conducted. In addition, information about patterns of purchase, particularly any seasonal or weekly variations, should also be considered when determining the time periods during which the interviewing will be conducted.

80. Questions relating to these latter points, namely the adequacy and representativeness of interview sites, arose in *Commerce Commission v Griffins Foods Ltd* (1997 DCR 797). Griffins argued that the mall intercept sites used did not ensure consumers who purchased potato chips from convenience stores or other outlets could also be surveyed. Boshier J accepted that *"a portion of the market has been ignored";* however, instead of identifying how this detracted from the survey, His Honour noted that "*the* [survey] *results must be seen accordingly".* Thus while it appears that deficiencies in addressing coverage error reduced the weight attached to the Commerce Commission's survey, the extent to which the survey findings were undermined by this criticism remains unclear.

81. This case highlights the need to ensure that the full range of purchase locations (or contexts where deception may occur) are considered. Thus, even where the relevant population may appear obvious, care must be taken to consider ensuring the sample represents consumers, or likely consumers, from the range of available purchase outlets.

4.1.4 Testing Coverage Error

82. The criteria outlined above noted the importance of ensuring opposing counsel has access to full details of the sample and selection procedures; failure to provide these details may lead to the rejection of the survey. Holland J made this point in *Noel Leeming Television v Noel's Appliance Centre Ltd* (1985) 1 TCLR 283 and acceptance of survey evidence turned on this issue in *ARA v Mutual Rental Cars (Auckland Airport) Ltd (*1987) 2 TCLR 141. In the latter case, the research company claimed providing details of the survey sample would breach respondents' confidentiality and refused to do so. The research company also failed to provide details of the instructions given to interviewers or the method of respondent selection.

83. The potential conflict between researchers' obligations to protect respondents' confidentiality and their need to provide full disclosure of the research undertaken must be carefully considered in the research design process. Survey interviews routinely collect personal information so that a proportion of the questionnaires can be audited; it would seen logical to include in the information provided at this point a statement that the details may also be used in legal proceedings. Respondents must, at that point, be given the option to withhold their details. This would not preclude providing their questionnaire to opposing counsel for review; however, if would mean that the questionnaire had no identifying information associated with it.

84. This approach would seem consistent with recent decisions. For example, in *Cookie Time Ltd v Griffins Foods Ltd* (2000: M1756/SW00), Glazebrook J held that failure to provide the names and contact details of survey respondents was "*not fatal*" (para 40); furthermore, His Honour accepted that the survey would have been conducted on a confidential basis. Glazebrook J noted that full details of the survey methodology had been produced and that the defendants had an opportunity to review the full set of questionnaires and, if they chose to, prepare and adduce their own survey evidence.

85. This case reinforces the need to ensure the survey questionnaires are available for critical scrutiny and appears to accept that this may take place without full access to details of the respondents who participated in the survey. The same issue arose in *Patience & Nicholson (NZ) Ltd v Cyclone Hardware Pty Ltd* (2001) 3 NZLR 490, where Cyclone argued that the survey evidence was inadmissible because the defendants did not have an opportunity to check the names on the customer list from which the sample was drawn, and because the actual names of the respondents had not been disclosed.

86. However, Hansen J held that the defendants had not been denied access to the information, and that the method used to select respondents would have resulted in an appropriate sample (para 92). Again, this case highlights the importance of providing information to opposing counsel that will enable them to check the validity of the estimates reported and appears to accept that this process may occur in the absence of full information about the individuals who participated in the survey.

87. In *ARA v Mutual Rental Cars (Auckland Airport) Ltd* (1987) 2 TCLR 141, Barker J was also critical of the research company's refusal to provide information that would have identified respondents participating in the survey. While he accepted that anonymity might be sought by respondents participating in surveys that examined sensitive topics, he did not consider the exploration of rental car companies *"were of a sensitive nature"* (1987 2 TCLR 153). His Honour noted that respondents could be advised the survey was to

be adduced in litigation and provided with an opportunity to withhold any information that would have led to their identification.

88. The need to ensure details provided by survey respondents can be tested, if necessary by the appearance of respondents in court, has been recognised in subsequent cases where full details of participants has been routinely provided, and where a sub-sample of respondents has been available for cross-examination. However, although details of individual respondents may not need to be disclosed, it is clearly vital to supply copies of the actual questionnaires, so that opposing counsel and experts assisting the defendants can test the basis of survey evidence adduced. In *Commerce Commission v Griffins Foods Ltd* (1997) DCR 797, the original questionnaires used to test consumers' understanding of the word "Slims" when used in connection with potato chips were destroyed and so could not be examined by Griffins. Boshier J agreed that "*A thorough examination of those forms might have revealed some issues as to credibility*", and noted that Griffins were unable to subject the questionnaires to the level of scrutiny ordinarily applied in cases such as this.

89. As well as providing information that enables opposing counsel to review the data collection process and to test the estimates reported from the data, it is also important to provide information about how the interviewing was conducted. This information should contain details of how the population was defined, how the sample was selected from within this, and quality assurance processes put in place to ensure the guidelines provided were followed. Failure to provide this information has attracted criticism and weakened the status of the survey evidence.

90. For example, in *ARA v Mutual Rental Cars (Auckland Airport) Ltd* (1987) 2 TCLR 141, Barker J described the devolution of responsibility from the survey designer to fieldwork staff, none of whom prepared an affidavit documenting the steps they took in selecting the organisations with whose staff they conducted interviews. His Honour referred to guidelines applied in Australian and UK decisions and concluded that the fieldwork supervisor should have prepared an affidavit outlining the instructions provided to interviewers and how respondents were selected. Evidence from the overall project manager about what he instructed the supervisor to tell interviewers was deemed inadmissible.

91. Barker J noted that the lack of information about respondent selection was not fatal, although its absence would have reduced the weight he placed upon the survey findings, had he found these acceptable overall. However, the failure to provide these details together with the absence of information about the instructions given to interviewers led him to conclude that the survey had not passed "*first base*" (1987 2 TCLR 155).

4.1.5 Summary

92. Addressing coverage error requires careful identification of the relevant population or populations of interest, and careful selection of a sample from within these. At the most fundamental level, researchers should consider the range of purchasers likely to be affected by allegedly deceptive behaviour (or likely to recognise a distinctive stimulus), and the patterns evident in their purchase behaviours. We suggest some more specific principles in Appendix 3.

4.2 Sampling Error

93. Consideration of coverage error leads logically to analysis of sampling error, an issue that has received more detailed attention than any other form of error. Unlike the other errors affecting survey research, sampling error can be quantified and, perhaps for this reason, legal counsel have paid particular attention to the precision of survey estimates.

94. When reporting survey evidence, many researchers make much of the error margin around each estimate. The error margin is a measure of the precision of the estimates as it provides a range within which the true value is likely to fall. The most commonly reported error margin is the 95% confidence interval; this error margin means that on 19 out of 20 occasions (95% of the time), researchers can be confident that the true population value falls within a prescribed range. However, it goes without saying that on 1 in 20 occasions, the estimate will not fall within the prescribed range.

95. Sampling error depends on the size of the sample (or sub-sample) analysed and the sampling procedure employed. The larger the sample size, the smaller the error margin and the more precise the resulting estimates. For example, the maximum error margin for an estimate based on a random sample of 1000 individuals is 3.1%, while it is 4.9% for an estimate of 50% based on a sample of 400.

96. However, while larger samples provide more precise estimates, the cost of obtaining a sample of 1000 is at between two and three times more expensive than the cost of interviewing sample of 400 individuals, while the maximum increase in precision is 1.8%. As a result, lawyers, their clients and their research teams must consider carefully the trade off between cost, sample size and precision.

97. This trade off must also be considered when the sampling procedure is determined, since this will also determine the precision of the estimates. Error margins (or confidence intervals) assume that estimates are based on a simple random sample, where every individual has a known and equal chance of selection. While some survey methods, such as telephone interviewing, achieve simple random samples, others, such as face-to-face interviewing methods, do not. For example, cluster samples are typically used in at-home face-to-face interviews because they reduce the fieldwork costs; however, depending on the size of the clusters, they may increase the error margins considerably. Face-to-face surveys conducted in shopping malls use a systematic selection procedure which, although technically not a simple random sample, is nonetheless assumed to be equivalent to one.

98. As well as considering the costs likely to be incurred obtaining different sized samples, lawyers must also consider the courts' response to survey samples and how the sufficiency of these has been assessed. We consider the evidence relating to this question in the section below, before examining the effects of the sampling procedure used.

4.2.1 The Courts' Assessment of Sample Size

99. Although most consumer surveys are based on a sample of several hundred respondents, some have been rejected because the judge considered the sample size inadequate. In *Pitstop Exhaust Ltd v Alan Jones Pit Stop International Ltd* (1987) 2 TCLR 502 the survey evidence was criticised as "*a survey of only 300 people*" and certain percentages were rejected as meaningless because they were based on very small cell sizes.

100. More recently, in *Automobile Club de L'Ouest, Aco v South Pacific Tyres New Zealand Limited* CIV 2005 485 248, the survey adduced by the plaintiff was criticised because *"The 273 people surveyed were an insufficient sample to provide statistically significant results"* (para 13). Similar comments were made more than a decade earlier in *Granny May's Management Pty Ltd v Whitcoulls Group Ltd* (1992) 5 TCLR 148, where Hillyer J noted, among other things, that the consumer survey was *"conducted with only 200 people".*

101. Similarly, in *Market Milk Federation of New Zealand Inc v Woolworths (New Zealand) Ltd* (1992) 4 TCLR 619, Anderson J noted that both the method used and the sample size made it "*inappropriate for [him] to attach any relevant weight to the particular evidence*" (1992) 4 TCLR 623.

102. These cases suggest that the sample sizes on which the estimates were based were insufficient to support the analyses the researchers undertook (hence the criticisms of the small cell sizes). For researchers, this suggests that, where the issue concerns the general population, small samples are not viewed favourably, particularly if the estimates required are based not on the total sample, but on a sub-sample drawn from this. Evidence that small sample sizes may lead to the rejection of a survey also comes from Folsom and Teply (1998), who observed that small scale surveys often did not meet with the court's approval.

103. Given these decisions, it was arguably predictable that the results of a pilot study of 25 individuals were rejected on the grounds that the sample was inadequate in *Yves St Laurent Parfums v Louden Cosmetics* (1997) 39 IPR 11. Interestingly, the courts' response to small samples does not appear to have informed survey development. For example, in a recent IPONZ decision, Case No. T15/2005, Assistant Commissioner Hastie found evidence submitted by three consumers was "*unqualified, similarly worded opinion*

evidence. As such it can not be relied upon as independent and probative. At best, it establishes only that three Aucklanders knew of the existence of the IMAX theatre in the Queen Street cinema complex" (p. 7). Williams J expanded on these concerns in *Imax Corporation v Village Roadshow Corporation Limited* (CIV. 2005-404-3248) where he discussed the evidence provided to IPONZ:

> *"In relation to this discussion, the evidence from the three members of the public is of little assistance in demonstrating awareness of IMAX's reputation and marks in the entertainment industry throughout the country. There were, after all, only three of them. The evidence does not disclose how the survey was conducted. It does not even disclose how the members of the public were selected, what questions were asked – though they seem to have been of a leading nature – the circumstances of the interview, or any of the other material which must be complied with for public survey evidence to be admissible."* (para 64).

104. While it is obvious that the IMAX survey suffered from several limitations, the reliance on the opinions of three consumers was clearly a major defect. Williams J's comments suggest that reliance upon evidence from selected consumers is unlikely to prove persuasive, and may be rejected entirely when no information about the methodology or Quality Assurance procedures employed is provided.

105. However, although these cases suggest small samples will almost invariably be criticised and, if not rejected, at least reduced in weight, some IPONZ hearings have admitted surveys based on a very small number of consumers. For example, in T29/2003, where Philips Electronics NV applied to register a 3-headed shaver shape, a survey of seventeen individuals was admitted.

106. Respondents were shown a picture of a 3-headed shaver and were then asked who made the shaver; eleven respondents identified Philips as the source. Assistant Commissioner Frankel found the survey admissible, although noted she did not necessarily find it compelling. However, she did not comment on whether she found the sample size, the respondent selection procedures, the representativeness of the sample, or the use of the staff employed by the applicant's attorney as interviewers, problematic. Instead, she commented on the logic of the applicant's argument, which she found wanting. It may be that limitations in the survey sample and data collection procedures reduced the credibility of the arguments she evaluated, although the decision did not make these points directly.

107. Where the population of interest is more specific, smaller samples have been more acceptable. For example, the opponent to the ROBOCUP application (T51/2002), Orion Pictures Corporation, adduced a survey of eighteen video store owners that had been conducted by a staff member of the opponent's patent attorney. Although this number appears very small, the total number of video stores in New Zealand is not likely to exceed a thousand, and the apparently small number of respondents was not criticised. Other issues, particularly the survey's failure to demonstrate consumers' awareness of the "ROBOCOP" movie, appear to have been more important in determining the weight given to the survey (p. 16).

108. Similarly, where the relevant population is trade professionals, smaller samples may be considered acceptable. Part of the evidence adduced in T29/2002, where Multichem Laboratorie's application to register the mark i-Profen was successfully opposed by Knoll AG, included a survey conducted by a member of Knoll Ag's legal team. The questionnaire was administered by telephone to 32 pharmacists or technicians assisting in pharmacies and examined whether respondents were aware of i-Profen. If they answered in the affirmative, they were asked how they became aware of this drug, whether the name suggested anything about the drug, and why they provided the answer they gave in response to the previous question.

109. Even though the mark i-Profen is not yet in use, four respondents claimed to be aware of it. Thirty associated i-Profen with Ibuprofen or thought it was similar to Ibuprofen, and twenty-five thought the name i-Profen sounded similar to Ibuprofen. Two respondents commented on the confusing similarity between the two names and their impression that many people confuse names such as these. The hearing officer considered the survey evidence fair and reliable as respondents were clearly identified, qualified to be included in the sample, the responses were clearly recorded, and respondents had not been led into a particular pattern of responses. However, although he noted that the survey evidence *"was not independently obtained"* and made no further reference to it, he did not comment on the sample size. It is possible that, because the results were so strongly suggestive of likely confusion, the number of individuals surveyed became less important.

110. Although few cases have made direct comments about the sample size, some generalisations are possible. First, the appropriate size of a survey sample depends on the size of the total population, thus small samples will not be rejected out of hand, if other aspects of the survey suggest it was competently conducted. Where specific trade audiences are the relevant population, a small sample might be quite understandable; however, although there is some evidence that small end user samples have been

accepted, it would seem prudent to consider 300, or even 400, as a minimum sample size for surveys involving the general public.

111. The actual size will depend on the level of analysis planned, as detailed breakdowns by user group or demographic characteristics require larger sample sizes. To some extent, the clarity of the findings may also be important, since if a high proportion of the population is likely to be confused, this should be apparent from a small sample. However, the prevalence, or likely prevalence, of confusion cannot be known in advance, thus relying on a small number of relevant consumers to produce clear evidence of confusion or deception would seem a risky strategy.

4.2.2 Assessing the Sampling Procedure Employed

Quota Samples

112. The second factor that contributes to sampling error is the sampling procedure employed. As noted in Section 4.2, the sampling errors normally reported for surveys are calculated for a simple random sample, where every member of a population has a known and non-zero probability of being selected to participate in the survey. In practice, however, the sampling procedures used often employ quotas, which ensure the sample fits pre-specified criteria, or clustering, which reduces the fieldwork costs.

113. Quota samples employ specific parameters that the survey sample must meet; these are normally used to increase the likelihood that the sample will match the wider population on known characteristics. They differ from screening questions, discussed in Section 4.1.1, which ensure only members of the relevant population are interviewed, and instead result in a sample that is balanced according to particular traits.

114. Gender quotas are often imposed on face-to-face and telephone samples and the resulting sample usually has an even split between male and female respondents. Imposing gender quotas can avoid badly skewed samples, where older females, a higher proportion of whom live alone and are available to participate in surveys, can dominate the sample composition.

115. However, logically, quotas should only be imposed if the variable used as the basis of the quotas is related to the behaviour of interest. Thus although quotas typically employ age and gender criteria, there is often little evidence that age or gender has a distinct relationship with confusion or recognition of a particular mark. Careless use of quotas may even produce a sample that does not align with the population of interest.

116. Evidence of the need to apply some caution to the use of quota sampling was provided in *Anheuser Busch v Budejovicky Budvar National Corporation* (2001) 3 NZLR 666, outlined above. The survey adduced in this case was criticised because it produced a sample that was unlikely to mirror the population of interest. For reasons that were not explained, the researchers had applied a 50:50 gender quota to their sample. Doogue, J noted: "*the gender mix in the sample was unlikely to reflect the gender mix of prospective purchasers of the products of either AB [Anheuser Busch] or BB [Budejovicky Budvar]"* 3 NZLR at 692. This comment suggests that quotas should only be used where they will improve the match between the sample and the wider population on the key variables being estimated.

Clustered Samples

117. Clustered samples are used in at-home face-to-face surveys to reduce interviewers' travelling costs. Instead of interviewing households at random, interviewers approach a cluster of households that are located around a randomly selected starting point. Interviewers are provided with a selection protocol that they use to ensure their selection of properties to approach is systematic. The size of the cluster may vary from five to up to twenty households.

118. The net effect of both quotas and clustering is that the Design Factor is increased; this, in turn, increases the confidence interval around each estimate and so decreases its precision (Koerner, 1980). In cluster samples, the design factor can be as large as two, and if quotas have also been used, it may be even larger. An increase in the Design Factor increases the error margins associated with the estimates. Thus, while a sample of 1000 people selected at random would normally have a maximum error margin of slightly over 3%, this could increase to 6% if the survey had involved clustered sampling. Details of how error margins and the Design Effect are calculated are provided in Appendix 1.

119. Although Morgan (1990) noted that it is desirable to ensure each member of the relevant population has a known, non-zero, probability of being included in the sample, probability samples are rarely adduced. In part this is because there may be no sampling frames for the population of interest from which a random sample may be drawn. Thus, where membership of the population of interest is unknown, non-probability samples are a pragmatic alternative. The Courts have recognised the problems researchers may face in generating a true random sample and have regularly accepted non-probability samples such as mall-intercepts (Jacoby and Szybilli, 1995; Bottomley 2001).

120. However, although non-probability samples have been accepted, Gastwirth (2003) noted that researchers should use a range of screening questions that identify actual and potential users of the brand or product category in dispute. This process ensures that the sample contains members of the relevant population, even though the characteristics of sample members were not known to the researchers in advance.

121. Overall, the courts have made very few comments about the sampling procedures used in consumer surveys, although they have accepted arguments by expert witnesses that quotas may have been inappropriately employed. However, although the paucity of comment may suggest decisions about respondent selection are less important, failure to consider the effect of sampling procedures on the error margin will leave the survey vulnerable to technical criticisms. While, on their own, technical criticisms have not usually resulted in the rejection of survey evidence, the cumulative effect of coverage error and poorly reasoned sampling decisions may lead to this outcome.

122. Non-response error is the third technical error that researchers must attempt to reduce. Thus, in addition to identifying the relevant population and drawing an appropriate sample from this, researchers must also gain responses from as high a proportion of sample members as possible. Failure to do so reduces the response rate and increases the possibility that those who completed the survey differed in some relevant way from those who did not respond or who could not be contacted (Dillman 2000, p.11).

123. The Federal Judicial Centre *Reference Manual on Scientific Evidence* (1994) cautions: *"If the response rate drops below 50% the survey should be regarded with significant caution as a basis for precise quantitative statements about the population from which the sample was drawn"* (p. 239). This implies that researchers must ensure high response rates or, failing this, must ascertain the extent and direction of non-response error. The following section examines non-response error, how researchers might address this error, and how the courts have reacted to surveys affected by non-response.

4.3 Non-Response Error

124. Although sampling error can be quantified (and is expressed through error margins or confidence intervals), the effects of other errors are often more difficult to estimate. Non-response error occurs when those people initially contacted to participate in a survey are either unavailable or refuse to complete an interview. Where these individuals differ in a material sense from respondents, non-response error occurs. To address non-response error, researchers attempt to maximise the response rate they achieve. Well-trained and confident interviewers increase the probability that those approached will agree to participate, and use of extensive "call-backs" can also ensure people who are difficult to contact will be reached and have the opportunity to be included in the sample.

125. However, while high response rates reduce the likelihood that serious non-response error will bias the survey estimates, it is possible for surveys with high response rates to be affected by non-response error. Conversely, surveys with low response rates are not necessarily affected by non-response error (although the potential for this to occur is greater). Nevertheless, the higher the response rate, the lower the potential for non-response error and researchers should endeavour to achieve as high a response rate as possible.

126. Where individuals refuse to participate in a survey, it is often difficult to obtain any information about them that would enable researchers to compare the characteristics of non-respondents (or refusers) against those of respondents. In some cases, the sampling frame employed will contain a known characteristic against which respondents and non-respondents can be measured, but this is not always the case. Moreover, unless the traits used as the basis of any comparisons have a clear relationship to the variables at the heart of the legal question, the comparisons may be meaningless.

127. For example, researchers can often access details of respondents' demographic characteristics (normally their age and gender) from sampling frames. Comparison of non-respondents' and respondents' demographic traits would enable a general assessment of whether those who were interviewed differed from those who were not, but may not necessarily provide insights into the behaviour of interest. That is, unless the characteristics researchers can use to compare the two groups have a clear relationship with individuals' likelihood of being misled or deceived, there may be little point in investigating the relationships further. For these reasons, non-response error remains very difficult to quantify.

128. Interestingly, few New Zealand judgments make specific reference to response rates, even when these are extraordinarily low. The response rate in *Anheuser Busch v Budejovicky Budvar National Corporation* (2003) 1 NZLR 479 was disputed and it is clear that a range of response rate formulae may be used to calculate the overall survey response rate.

129. The American Association of Public Opinion Research has issued detailed guidelines regarding the presentation of survey results and disclosure of methodological information; adoption of these guidelines would help standardise the information reported and how this is presented. Details of these formulae are provided in Appendix 2.

130. We suggest that Response Rate 5 represents the most rigorous formula, since it assumes the proportion of ineligible respondents among those who were not contacted or interviewed is zero. If variations on this formula are used, the assumed proportion of ineligible respondents among the non-contact group should be clearly outlined.

131. Although response rates have not attracted detailed comment from judges or hearing officers, the trend towards declining response rates suggests this source of error should become increasingly important. For example, response rates to telephone surveys may be lower than 20%, and some face to face surveys do not achieve a response rate above 30%. According to the United States Federal Judicial Centre *Reference Manual on Scientific Evidence* (1994), response rates that fall to this level raise serious questions about the validity of the survey estimates.

132. Some of the most important tools researchers can use to combat falling response rates include the survey itself and the interviewers who administer this. A well-presented and engaging survey that is introduced by a competent and personable interviewer may be sufficient to elicit support from a reluctant respondent. The following section examines these points and the errors that arise from the design and administration of the questionnaire.

4.4 Measurement Error

133. Measurement error, the final type of error we consider, is arguably the most critical factor for researchers to consider and yet, ironically, tends to receive less attention than technical errors, such as sampling error. It includes error from a wide range of sources, including the design format and administration of the questions and other aspects of interviewers' behaviour. Dillman (2000) suggested measurement error occurs when "*a respondent's answer to a survey question is inaccurate, imprecise, or cannot be compared in any useful way to other respondents' answers"* (2000 p. 9).

134. Inaccuracy, or a lack of comparability, may arise for several reasons; Hudis (2000) explored these further when he examined criticisms the US courts have levelled at survey evidence. These criticisms, which highlight the importance of taking steps to minimise measurement error, include poor procedures (such as the use of leading questions) through to highly questionable practices (such as the exclusion of responses deemed unfavourable). Measurement error thus covers a range of methodological problems; the following section examines flaws in question design, and questionnaire construction and administration, and considers how these may affect the validity of respondents' answers.

4.4.1 Question Wording

Open-Ended Questions

135. Respondents' understanding of questions, and their ability to provide clear and accurate answers to these, depends heavily on the structure of the question itself. Care must be taken to avoid bias in the design and wording of the questions put to respondents, since this will skew the resulting response distribution. For this reason, many surveys use open-ended questions, since the absence of formal response options means respondents are free to provide whatever answers occur to them.

136. Open-ended questions do not present respondents with an array of options from which they choose one or more, but allow them to articulate a response in their own words. Because of their open format, this style of question reduces the risk that it will be considered leading, although it does not eliminate this risk altogether.

137. For example, the survey adduced in *Patience & Nicholson (NZ) Ltd v Cyclone Hardware Pty Ltd* (2001) 3 NZLR 490 questioned 181 stockists and end users of cutting tools; of these, 130 had heard of P&N and were asked: "*What does P&N mean to you?*". This very open question may have assumed that "P&N" would mean something to respondents, but it did not suggest what the meaning could be. A follow-up question

examined what "P&N" meant and revealed that 69% of the stockists and half the end-users associated "P&N" with Patience and Nicholson.

138. The question used in the P&N survey is non-leading insofar as it does not suggest to respondents any meanings they might associate with P&N. However, while the question wording was not criticised by counsel for Cyclone, it was suggested that repeated use of "P&N" in the questions was likely to heighten the salience of this brand. As a result, counsel argued that the emphasis placed on P&N in earlier questions would increase the likelihood that respondents would suggest this brand, rather than others, when they were asked which brands they stocked. Nevertheless, Hansen J held that the survey was admissible and that it provided *"evidence of a strong association between the letters P&N and Patience & Nicholson New Zealand among stockists of cutting tools… and also among end users of goods manufactured by the plaintiff"* (para 96). This finding was subsequently upheld on appeal, where Keith J found that the decision to accept the survey was *"open to the Judge on the evidence"* [2001] 3 NZLR 490, 530.

139. Use of an open-ended question with a tightly defined and knowledgeable population, as occurred in the P&N case, reduces the need to create a context for the question. However, where respondents are selected from the general public, and where their levels of knowledge are likely to vary, researchers create a more specific frame of reference within which they anchor their questions. While this process guides respondents to the legal issue of interest, care must be taken to ensure it does not lead them to favour a particular response.

140. This danger was not avoided altogether in *Automobile Club de L'Ouest, Aco v South Pacific Tyres New Zealand Limited* CIV 2005 485 248, where Wild J commented that the question used was "*not an entirely open one"* (para 51). The question in dispute asked respondents: *Thinking in terms of motoring what do these words* [Le Mans] *mean to you?* Although Wild J did not outline his concerns in detail, it is clear that use of the words "*in terms of motoring"* created a context that he believed respondents might not otherwise have considered. He levelled similar criticisms at the second two surveys adduced in this case, arguing that the creation of context that related to tyres or cars also moved sufficiently far from being a completely open question as to cause some concern. Wild J later summarised his concerns thus: *"The way in which the questions were cast was likely to have skewed the results away from the geographic associations some consumers may well have had with the words 'Le Mans'.* [para 55].

141. While completely open questions avoid allegations that respondents have been led, they are likely to elicit a wide range of responses that researchers must subsequently code and interpret. Thus while open-ended questions reduce the potential that

respondents are directed to a particular answer or view, this question format may present difficulties when the data are being prepared for analysis. The coding frame, or set of classifications used to group similar answers, can be strongly disputed by opposing counsel and experts, as it is possible to develop different classifications that may result in different estimates. Use of multiple coders and the ability to demonstrate a high level of agreement between these, may help researchers establish the robustness of the frame they have used. However, reliance on data from open-ended questions carries with it the risk that a successful challenge to the coding frame will lead to rejection of the analyses and conclusions based on these.

142. In addition, use of open-ended questions means interviewers must be carefully trained to record every comment, including the non-verbal gestures respondents must make (such as pauses, or "um"). When faced with the demands of managing their clipboard, maintaining some eye contact with respondents, and ensuring there are not long silences while they record a detailed response, many interviewers resort to summarising the answers respondents provide. While this is a pragmatic solution to managing the interview, it may nevertheless compromise the accuracy of the responses and create opportunities for opposing counsel and experts to question the validity of the resulting estimates.

143. Concerns about the accurate recording of survey responses arose in *Cookie Time Ltd v Griffins Foods Ltd* (2000, M1756/SW00), where counsel for Griffins noted that some questionnaires contained two different types of handwriting, a situation that would not normally occur if interviewers transcribed respondents' answers. In addition, the responses to some questions had been incorrectly recorded and there was evidence that some responses had been summarised rather than recorded verbatim. Although Glazebrook J did not reject the survey findings as irretrievably flawed, he noted that the results could be discounted, should the survey be adduced at trial.

144. The problems interviewers face in recording verbatim answers can be avoided through the use of tape-recorders. However, while these reduce the need to write very quickly, they may introduce other problems. First, while recording the interview reduces the burden on interviewers, tape recorders are necessarily obtrusive and some respondents may feel overly self-conscious when responding, particularly if the questionnaire explores sensitive topics. Thus while recording the interview may ensure that all the comments respondents make are recorded, this approach may limit the range of responses interviewees are willing to provide.

145. Second, if interviewers deviate in any way from the questionnaire, for example, if they ad lib comments, paraphrase instructions, fail to follow skip instructions, or deviate

in any way from the interview script, their errors become evidence that opposing counsel will use to undermine the validity of the survey. We examine this problem in more detail later in this section. To reduce the difficulties associated with open-ended questions, researchers may employ closed-questions or develop a context that respondents use to frame their answers; we discuss the merits of this approach in the following section.

Closed Questions

146. Because of the difficulties in recording and coding open ended questions, some researchers have used closed questions that employ pre-coded response options. These simplify the task respondents are required to complete, reduce the onus on interviewers to record faithfully every comment respondents make, and lessen the likelihood of disputes over the coding frames used. However, to function successfully, researchers must ensure that the coding categories presented to respondents are balanced and include the full range of answers likely to be provided. Questions that are fielded with incomplete coding categories may bias the resulting estimates, thus inviting criticism from opposing counsel. In addition, researchers using closed questions need to ensure that the presentation of the response options does not result in any primacy or recency effects that favour a particular response. To guard against this possibility, the response options are often rotated so that any order effects are randomised.

147. Both open and closed questions have been accepted by the US courts. In *Union Carbide Cop. V Ever-Ready Inc.* 531 F.2d 366 (7th Cir. 1976), open-ended questions were used to explore who respondents thought produced a product; what other products made by the same company were, and whether companies that put out a particular product would have permission of any other company to put out that product and, if so, why. By contrast, *Squirtco v Seven-Up Co* 628 F.2d 1086, 1091 (8th Cir. 1980) explored the legal question more directly using a closed question that presented balanced response options:

> "*Are these two products made by the same company, by different companies, or don't you have an opinion about that one way or the other?"*

148. However, use of closed questions is contrary to the decisions issued in some recent cases. For example, in a UK case, *Black & Decker v Atlas Copco Aktiebolag* (O-281-03), the court criticised the use of closed questions, which it accepted would result in an over-standardisation of answers. Instead, it was argued that it was preferable to collect the verbatim responses of respondents; Hearing Officer Landeau considered that this process would ensure the responses had not been amended to fit pre-defined coding categories.

149. Even where the questions used are closed, the interpretation of the findings can be disputed. In *Pitstop Exhaust Ltd v Alan Jones Pit Stop International Ltd* (1987) 2 TCLR 502, Wylie J criticised the lack of clear logic in the analyses presented to him. His Honour noted that of 300 people who had been interviewed, 9% (27 people) said they had been to PitStop. He went on to comment: *"Surprisingly, considering they had been to the premises, 25 percent of those (although what 25% of 27 people is I do not know) did not know what the business did. Other figures were given. Some simply do not make sense to me, eg 1.4 percent of 27 people knew of Pitstop from television advertising. On its face the percentage mentioned is absurd. What it all means in the present context I fail to understand"* (1987) 2 TCLR 508. These comments highlight the importance of ensuring the analyses presented can be easily linked to the total sample, and that the percentage figures and individual responses are also clearly and logically linked.

150. While it is true that pre-coded categories may limit the range of comments respondents make, it is worth noting that respondents often make a wide range of extraneous comments that have little or nothing to do with the questions put to them. Use of open-ended questions may therefore result in a considerable volume of irrelevant material and almost invariably increases the length of the interview. Closed questions avoid these problems, but encounter others if they do not present respondents with a full range of response options. Researchers faced with the need to choose between open and closed questions can rely on pre-testing to develop and review the relevant response options.

Pre-testing

151. Both Dillman (2000) and Morgan (1990) stressed the need for careful questionnaire design and suggested pre-testing to make certain that respondents understand the questions put to them. Pre-testing also ensures that the question wording does not favour particular responses and that the list of response options provided in closed questions is complete.

152. However, although pre-testing of survey questionnaires is sometimes conducted, pre-tests normally often only check the mechanics of the questionnaire, such as its flow, and whether the skip instructions function correctly. Only rarely does pre-testing examine issues of validity, such as whether respondents and researchers share a common understanding of the survey question. Belson's double-back is a well-established technique specifically designed to examine respondents' interpretation of survey questions (Belson, 1981). This procedure involves administering a question to respondents, and then reviewing respondents' interpretation of the question. By asking respondents what question they were answering, the technique enables identification of mis-understood terms and intentions. In particular, researchers may use the technique to

investigate whether the wording used in the questions parallels the wording respondents use to describe the issue at hand. Other techniques include cognitive laboratories, where respondents think through their answers aloud, thus enabling researchers to observe how the survey questions are interpreted. Questions that are critical to the legal issue of interest should be carefully pre-tested to ensure the questions respondents believe they are answering are in fact those that researchers intended them to address.

153. Another approach to quality assurance testing could use a conference where an independent expert and legal counsel representing the parties in the dispute review the survey. Given the competing interests represented, this process would seem likely to elicit all the response options required. In addition, it could reduce disagreements over the need to create coding classifications, since the response options would be pre-coded. In principle, agreement on the questions used should enable more detailed attention to be paid to interpreting the estimates themselves, and would shift attention away from the process used to produce these.

154. Challenges to questions that arguably did not include all the possible responses have been made in some cases. For example, in *Cookie Time Ltd v Griffins Foods Ltd* (2000: M1756/SW00) the survey assumed the only logical responses were ones that related either to the plaintiff or to the defendant, and that corresponded to the different distribution channels utilised by Griffins and Cookie Time. However, some responses did not fit neatly into these mutually exclusive categories, leading counsel for Griffins to suggest that other brands may have been available, and responsible for the alleged confusion. In the absence of evidence that another brand had been available, Glazebrook J did not appear to place any weight on this suggestion, although he noted that the conclusions drawn from the survey may not have been the only conclusions available to the researchers.

155. Irrespective of the type of question used, interviewers must exercise careful judgment to ensure they have elicited the full range of responses their interviewee wishes to provide, while at the same time not pressuring them to guess, simply in order to provide a response. This latter issue may become problematic if respondents' knowledge of the survey topic is incomplete. For this reason, Morgan (1990) recommended that respondents should not be asked to complete tasks that do not correspond to the behaviour of interest and noted that they should be instructed not to guess when answering questions. Problems such as guessing by respondents are difficult to prevent, although careful training of interviewers and question wording that explicitly includes a "don't know" or "no opinion" response can reduce the scale of this potential problem. As noted above, another approach involves the creation of a context that limits the likelihood respondents will provide completely irrelevant answers. We explore how researchers may achieve this goal in the following section.

Survey Context

156. Many judges have noted the importance of ensuring respondents are not asked to embark on fields of speculation that would not otherwise have occurred to them, had the question not been put to them (see *Anheuser Busch v Budejovicky Budvar* [2001] 3 NZLR 666). It is also important that respondents can answer the questions put to them, and that they are able to use their knowledge and experience to provide informed and accurate responses. However, where questions lack context, they may, ironically, confuse consumers, even if they do not influence their views in any way. Researchers must thus strike a balance between leading respondents on the one hand, and avoiding questions that are so open-ended respondents are uncertain what they are being asked. In addition, they must ensure the survey has external validity.

157. This latter issue has received attention from both academic researchers and the judiciary. Eko (1998) and Judge Posner of the 7th Circuit, noted earlier, both highlighted the importance of the scenario within which confusion is likely to occur, and ensuring that the survey design replicates or simulates this. Failure to create a research context that closely parallels the way consumers make purchase decisions, or the setting in which they are likely to encounter a trademark, reduces the overall validity of the survey. As McCarthy (1998) noted, "*the closer the survey methods mirror the situation in which the ordinary person would encounter the trademark, the greater the evidentiary weight of the survey results*" (p.237).

158. Surveys that do not reflect the types of behaviour consumers would normally engage in, or that limit the range of behaviours in which they can engage, are therefore likely to have little weight attached to them. It is now widely accepted that the tasks respondents perform in a survey should correspond to the legal question of interest and explore or simulate a realistic behaviour (Kearney & Mitchell, 2001). When evaluating the merits of consumer survey evidence, judges may therefore consider how the disputed marks are used in the relevant marketplace, and the extent to which the survey provides insights into that marketplace.

159. Surveys have not always performed well when compared against this criterion. For example, Sheppard, J. in the Federal Court of Australia decision of *Interlego AG & Anor v Croner Trading Pty Ltd* (1991) 102 ALR 379 noted that he could not place weight on a survey that involved: "*a hypothetical situation which was… so artificial as to make it quite a dangerous guide to…reactions of actual shoppers."*

160. Similar concerns were raised in a later New Zealand decision. In *Wineworths Group Ltd v Comité Interprofessionel du Vin de Champagne*(1992) 2 NZLR 327, Cooke P noted:

"It is necessary then to focus upon the significance of the name 'Champagne' in the marketplace, how it is used and how it is understood in the course of trade. That is of particular importance in this case because some of the evidence, especially of the English language experts and the public opinion survey experts, did not always identify the context or circumstances in which they found the name to be used in certain ways" 337 (1992) 2 NZLR 327. In this particular case, His Honour found that careless use of a word, such as "*champagne*", by the public did not necessarily invalidate the distinctiveness of the brand name when used by members of the trade.

161. Failure to establish a purchase-related context was also apparent in the survey adduced in *Anheuser Busch v Budejovicky Budvar* [2001] 3 NZLR 666. This survey examined confusion over the word "Budweiser" and asked respondents to examine three bottles of imported beer. After some intervening questions, they were asked to recall the bottles they had seen. This test did not explore whether respondents confused the two beer brands, Budejovicky Budvar and Budweiser, nor whether they would be likely to do so in a normal purchase situation. Rather, it examined the extent to which respondents could recall foreign brand names, at least some of which they may not have previously encountered.

162. As a result, the Budweiser survey did not approximate a purchase situation consumers might encounter in a retail environment and the measures of recall obtained were thus not equivalent to measures of likely or actual confusion. Doogue, J commented: "*of more importance to me is the fact that the interviewees in the market survey were not being faced with a market situation and that the degree of confusion, even in an artificial situation, was slight. It seemed to me to be clear that, faced with a market situation, the degree of confusion had to be less…The market survey simply did not address the reality of a market situation."* 685 [2001] 3 NZLR 666.

163. His Honour went on to comment: "*the survey was as far removed from a practical purchasing situation as the presence of a number of bottles before me in the courtroom. It had nothing to do with the likelihood of imperfect recollection of 'Bud' or 'Budweiser' marks or AB's product when trawling the shelves of a supermarket or a liquor wholesaler or the like and being confronted not a by a six-pack of 'Budweiser' but by a four-pack of 'Budejovicky Budvar… The survey was conducted in an artificial and unrealistic context simply too far removed from that in which actual consumer decision making would be made."* 692 [2001] 3 NZLR 666.

164. The need to ensure respondents' task paralleled the situation they would encounter in a retail store also received attention in an earlier case, *Bluebird Foods Ltd v Cerebos Greggs Ltd (1998)* (CP323/98). The plaintiff, Bluebird, sought an injunction to prevent

Cerebos Greggs Ltd from marketing a corrugated snack food using the name "Mexican Waves" or "Waves". They argued use of these names in association with a grooved savoury snack food was deceptively similar to their product "Grainwaves", and would mislead consumers. Survey evidence was adduced to test consumers' recognition of a savoury snack product; this found that 85% of those surveyed claimed to recognise the "Mexican Waves" product when this was shown to them. However, 71% of this group named the product as "Grainwaves" and a further 5% indicated some aspect of the "Grainwaves" name in their response. Of the 71% who named the product "Grainwaves", three quarters used the shape of the product to support their decision. Counsel for Bluebird submitted these results suggested that more than half the market confused Mexican Waves with Grainwaves.

165. Interestingly, the discussion over this survey evidence centred not on the questions used or the sample employed, but on the fact that respondents were shown the product without packaging, in what Cerebos Greggs' counsel described as a "post-sale" situation. Cerebos Greggs responded that confusion had not been established in a point of sale situation, when both products would be shown in their respective packaging, which made the different brand names clear. However, Smellie J was guided by *Levi Strauss & Co v Kimbyr Investments Ltd* (1993) 28 IPR 249 where Williams J had held that post-sale confusion was relevant.

166. Similar concerns were raised in *Cookie Time Ltd v Griffins Foods Ltd* (2000: M1756/SW00), where the parties were involved in a dispute over the packaging used to present novelty biscuit items. Respondents were shown plastic buckets containing small biscuits; however, Griffins' counsel argued that respondents were not given an opportunity to read the label on the bucket, that it was rotated while they were looking at it, and that it was shown at waist height (rather than at eye level). This combination of factors allegedly created a situation unlike a normal shopping experience, where consumers would have the opportunity to examine a product prior to purchase.

167. These cases highlight the need to create a context that has external validity. While the types of questions used play a key role in shaping the context, the choice of survey mode may also affect researchers' ability to present respondents with realistic tasks and choices. That is, the survey mode itself can become problematic if it does not allow for the creation of an environment where respondents would normally expect to encounter the product or mark under investigation. For example, in Frucor Beverages Limited's (TM 30/2003) application to register the colour green as it related to various beverages, evidence attesting to the distinctiveness of this colour was based on a telephone survey. Assistant Commissioner Brown QC noted "*I was a little surprised that a survey to determine the level of distinctiveness of a colour trade mark was undertaken by means*

of a telephone survey." He later noted that he did not find the survey *"particularly compelling"*, although other external market data documenting the use of the colour green in the marketplace held more force.

168. For survey researchers, these cases present interesting design issues. While the external validity of the study would be increased if the interviewing took place in an actual retail context, the range of other variables likely to be present could reduce the internal validity, as not all of these would normally be able to be controlled. For example, respondents in a retail store could be affected by in-store announcements, end-of-aisle displays, and brand juxtaposition; none of these variables would be under the researchers' direct control.

169. Researchers' response to the need for external validity has typically been to design a questionnaire that creates a purchase context for respondents. Although not physically present in that context, visual stimuli, or even the wording used in questions, can enable respondents to create vicarious contexts and to locate their responses within these. Thus, while not measures of actual behaviour, experimental settings can, to some extent at least, simulate purchase contexts. More importantly, an experimental design allows researchers to exert tighter control over the range of products and brand attributes respondents consider.

170. Having decided what types of questions to put to respondents, and how to frame these, researchers must also test their questions to ensure these are fair, balanced and non-leading. These criteria are particularly important, since questions that direct respondents to a particular answer, or that lead them through a questionnaire such that their views are shaped in a particular way, will inevitably attract criticism and may lead to the rejection of the survey. The following section discussions these issues in more detail.

Leading Questions

171. Many researchers have identified leading questions as problematic and a reason why judges have rejected survey evidence. Kearney & Mitchell (2001), for example, cautioned against questions that invite speculation on particular topics or that lead respondents in a certain direction. By this, they mean questions that introduce concepts or points of comparison that respondents may not have considered.

172. However, while the arguments against leading questions seem clear, designing surveys that do not lead respondents, at least to some extent, may be difficult. As the previous section illustrated, failure to provide a context may result in irrelevant responses. On the other hand, creation of a context that has little in common with a

purchase situation may detract from the survey's external validity. Thus, despite the fact that researchers may wish to test the association between two brands or marks, or explore the extent to which a claim may mislead consumers, they must approach these questions obliquely. Questions that outline the relationship under investigation almost inevitably become leading.

173. Evidence of this problem was apparent in *Anheuser Busch v Budejovicky Budvar* [2001] 3 NZLR 666, where the critical question used to explore brand confusion asked respondents whether they thought BB was in some way associated with AB. A number of the respondents to whom this question was put agreed that an association existed. However, it is not clear whether they made this association independently of the question, or whether they answered in the affirmative because they did not expect to be asked a question about an association if in fact no association existed. Doogue J commented: "*I was particularly concerned with the use of a clear leading question, question 5, in the course of the survey.…If there was an allowance for the misuse of the leading question, the level of apparent association between BB's bottle in the artificial situation and AB's name would have reduced significantly"* (692 [2001] 3 NZLR 666).

174. Concerns over leading questions in consumer surveys have arisen in several cases. Fisher J discussed a problematic question used in a survey adduced in *Cerebos Greggs Ltd v Unilever New Zealand Ltd* (1994) 5 NZBLC 103, 497. This case examined a dispute over the type of coffee used in coffee bags and whether the descriptions used were misleading. A consumer survey asked respondents: "*What type of coffee would you expect to be inside the coffee bags?";* Fisher J commented that "*Unless the respondents were prepared to challenge the form of the question, it tended to push them into the assumption that there was only one type in the bag and to suppress any tentative thoughts that there might be two, namely roast and ground and instant."* Although His Honour found that the question could have been more appropriately phrased, he accepted the survey, but accorded less weight to it.

175. In *Commerce Commission v Griffins Foods Ltd* [1997] DCR 797, Boshier J found that the question wording would have shaped respondents' answers to the survey questions. The first question asked respondents if there was *"anything said or shown"* that led them *"to feel that the product is 'special' or 'different' in any other way to other potato chips made by ETA or other companies?"* The second question went on to consider whether respondents *"Would …expect there to be any more or any less of any ingredient in this packet of Slims chips?"* Boshier J considered that this second question prompted respondents to consider issues that might not otherwise have occurred to them; this flaw was one of a number that led him to reduce the weight attached to the survey.

176. However, the question used in Frucor Beverages Limited's application (T30/2003) to register the colour green, which asked respondents *"Do you associate bright lime green packaging with any particular energy drink?"*, was not criticised as a leading question. Although the structure of this question appears to introduce an association between the colour green and energy drinks, it does not suggest a particular brand to respondents. Thus the question still allows respondents should arrive at an association independently, should they associate the colour green with specific beverages.

177. The design of the question used in *Levi Strauss Co. v Kimbyr Investments Ltd* (1994) 1 NZLR 332 also met with approval. This case involved a dispute over the use of a tab on the pocket of jeans, a feature Levi Strauss argued was so distinctively associated with their design that use of a tab in any position would confuse and mislead respondents. To test this claim, respondents were shown a drawing of a pair of jeans that featured a tab on the back pocket in the location that Kimbyr used. Of the 500 people interviewed, half thought the brand of jeans in the picture was Levis, even though the tab was not placed in the correct position for Levis' jeans. Only 16% recognised the jeans were a brand other than Levis.

178. Counsel for Kimbyr criticised the survey questions, arguing that the initial brand awareness question had no intrinsic purpose. However, Williams J accepted that the purpose of the question was *"to channel the response to a subsequent question."* Williams J quoted from Whitford J at p 379 of *Levi Strauss v Shah*, where the latter noted:

> *"I have said very little about surveys, because in part at the very best they are only of marginal significance and in part because these surveys, and like all surveys in legal proceedings, are open to a good deal of criticism. They in fact illustrate the difficulties faced by anybody trying to conduct a survey. Creditably enough, those conducting polls for the plaintiffs started off by asking a question that can be said to be wholly unobjectionable, in that it was not leading, and the only result of that was they did not really get any sufficient answers directed to the point at issue to be statistically of any significance."*

Williams J recognised that surveys need to follow a sequence of questions that bring respondents to the issue at the heart of the dispute and found that the question to which Kimbyr objected was neither a leading question nor framed in such a way that it would detract from the value of answers given to subsequent questions.

179. Such an approach has been used successfully in Cadbury's registration of the colour purple (CT 32/2003) and was also accepted without criticism in Effem's application to

register the colour orange (CT43/2003). The latter survey provided respondents with an unbranded piece of packaging and asked if they associated this with anything (CT43/2003). The questions put to respondents became increasingly specific until they were asked whether they associated the piece of packaging with any rice brands, at which point the connection between the colour and a particular brand could be noted.

180. Yet, although Effem's approach followed the guidance provided in *Levi Strauss Co. v Kimbyr Investments Ltd* (1994) 1 NZLR 332, and the sequence of questions was carefully designed not to lead respondents, the results were not unambiguous. Opposing counsel used the low levels of association elicited in response to the very general questions to argue that an association between the colour and the brand did not exist (since, if it had, they claimed, it would have been apparent from the outset).

181. Thus, while this discussion highlights the acceptability of moving from a very general to a more specific context as an appropriate mans of bringing respondents to the issue of interest, it also illustrates the dangers implicit in very general questions. We discuss the use of funnelling questions and question order effects in the following section.

Question Order and Focus

182. The preceding sections suggest researchers must consider the types of questions put to respondents and the response options associated with each of these, as well as the order in which they present questions to respondents. Gastwirth (2003) summarised the approach endorsed by the courts when he suggested that researchers could minimise their vulnerability to criticism by ensuring the questions start by exploring general issues before moving to examine the specific issues at dispute. In addition, he suggested rotating the order of response options used in any closed questions, to randomise order effects and the chances that some options may be favoured because of their position in a list.

183. Although Williams J, in *Levi Strauss Co. v Kimbyr Investments Ltd* (1994) 1 NZLR 332 accepted that surveys comprise a sequence of questions that move respondents from general introductory questions to specific questions that examine the issues of interest, they must still avoid leading respondents to particular conclusions. Thus, questions must be framed neutrally so a specific response is not signalled as more appropriate, and they must move fairly and logically from more general to more specific issues.

184. As well as following an appropriate route to the legal question of interest, the survey must also define and explore this correctly, as even well-designed questions may carry little weight if they do not address the legal question of interest. Thus, providing evidence that the estimates are reliable, non-leading, directed to the appropriate

sample, and carefully analysed and interpreted, will not compensate for questions that do not address the matter at issue.

185. Researchers' failure to consider the legal issue correctly has undermined the usefulness of survey evidence. For example, Gault J, in *Allied Liquor Merchants Lit v Independent Liquor (NZ) Ltd* (1989) 3 TCLR 328 noted that the surveys he had been asked to consider did not ask *"the real question in issue, namely, 'What did the purchaser think he or she was buying?'… While the survey indicates that to some extent many purchasers may not care, I do not think that is of any particular relevance."*

186. Very similar criticisms were raised by Dr Barton, counsel for *International Cellars in Blenhaven; International Cellars (Marlborough) Ltd v Montana Wines Ltd* (1989) 3 TCLR 115. He argued that the survey adduced in this case asked respondents whether the name Blenhaven was connected with another brand; this was not, he submitted, a question capable of testing whether respondents confused those brands.

187. Although not raised directly in the Assistant Commissioner's decision, the use of a showcard with the word "Blenhaven" was criticised for displaying an artificial stimulus that did not parallel the purchase context in which respondents would encounter this word. Thus, as well as failing to address the legal question of interest, it could also have been argued that the experimental context lacked external validity.

188. However, Assistant Commissioner Martin was more sympathetic to the survey and, instead of considering whether the questions posed addressed the legal issue, he commented that direct questions about confusion could not have been put to respondents. While the Assistant Commissioner's assessment on this latter point is correct, his comments raise a different issue to the criticism Dr Barton levelled at the survey, which suggested that the actual question used was inappropriate.

189. Dr Barton pursued this latter point in the third argument he raised about the survey evidence. Here, he suggested that because the survey tested brand association, the resulting estimates could not be interpreted as indicating brand confusion. Thus, although 53.4% of those surveyed associated "Blenhaven" with another wine, and, of these 94.6% associated "Blenhaven" with "Blenheimer", Dr Barton argued that these estimates did not offer insights into the proportion of the market that might be confused.

190. Assistant Commissioner Martin accepted these arguments, noting that: *"I consider … the response given by the interviewees to Q2 after they had seen the card with the word 'Blenhaven' printed upon it is not conclusive as to the likelihood of confusion… probably the connection that they refer to takes the matter little further than simply*

acknowledging that they know of a wine sold under a trade mark that commences with the letters 'Blenh'. The question of likelihood of confusion between the marks as a whole is a separate matter." (1989) 3 TCLR 125.

191. Developing questions that directly address allegations of confusion without leading respondents also proved contentious in *Commerce Commission v Griffins Foods Ltd* [1997] DCR 797. Boshier J noted that the question purporting to examine deception did not do so directly. The question concerned asked respondents if the brand name *SIims* was "*just a brand name for the product"* or whether it was *"a brand name for the product which tells you something about the product".* On the basis of responses to this question, the research company and its expert concluded that around 30% of respondents thought there was a "*fat/health benefit meaning"* and, extrapolating from this, that a similar proportion *"believed that the SIims brand would be a better choice for people wishing to avoid putting on weight".* However, Boshier J did not accept these conclusions and found that the survey did not test whether purchasers of *SIims* had been misled, but rather what their immediate reaction to a packet of S*lims* potato chips was. Thus while he accepted that some of the respondents could have been misled by the packet, he did not accept that the survey evidence enabled him to determine the proportion affected in this way with *"any degree of exactitude".*

192. In trademarks disputes, survey evidence indicating the awareness of a brand or mark in a market is sometimes provided. However, evidence relating to awareness of a mark is not necessarily evidence that the mark has become distinctive. In IPONZ case 2001/16 TM No. 271895, which considered an application to register a range of marks that included the letters PH, a survey adduced by the opponent investigated brand usage. Of the 83 responses obtained, 40 indicated that the respondent regularly used Weldwell, while 14 indicated the respondent regularly used Philips (the opponent) and 2 named PH as the brand they used regularly. Although Assistant Commissioner Howie did not explain his concerns in detail, he noted that he did not find survey evidence of this type of be particularly helpful (p. 16). It seems likely that he found the evidence did not address the question he was asked to consider; that is, he was charged not with determining which brand respondents used regularly, but how they perceived the disputed marks.

193. Similar concerns about the direct relevance of the evidence adduced were raised in the ROBOCOP decision. Respondents were administered four questions: whether they had any ROBOCOP movies; if so, which ones they had, how many copies of each movie they had, and when the store had received its copies of each movie. Seventeen of the eighteen stores had a ROBOCOP movie or game that had been acquired before the application date.

194. The Hearing Officer noted that this evidence was "*helpful to the extent of establishing an awareness of the ROBOCOP movies, series and/or game amongst those persons involved in the video rental industry*". However, the survey as reported did not test awareness of the movies, but availability of the movies through particular retail outlets. Nor did the opponents provide evidence of the frequency with which the movies were hired, although Assistant Commissioner Walden surmised that because the stores included in the sample had comparatively few copies of the different videos, it was unlikely to be a movie in high demand.

195. The Court of Appeal decision in *Mainland Products Ltd v Bonlac Foods (NZ) Ltd* (1998) 8 TCLR 224 also criticised the survey evidence relating to consumers' understanding of the word "*Vintage*". Gault J found that those who designed the questions did not appreciate the fundamental issue the court was required to address, namely: *"whether, when used on a label for a cheese product in the manner Bonlac has adopted, the word "Vintage" at the end of 1992 would have been taken as a description or as indicating a connection with a particular supplier (whether or not identified).*

196. The survey involved a telephone administered questionnaire that contained five questions. The first asked respondents to identify all the types of cheese they could think of, the second asked them if they had heard of vintage cheese and then asked them what "Vintage" cheese was. The fourth asked what brand or brands of vintage cheese they were aware of while the final question asked whether respondents thought the word "Vintage", when used in relation with cheese, was a type of cheese, a brand of cheese, both a type and a brand, or neither a type nor a brand. The final question also included a "Don't know" response option ([1998] 8 TCLR 232).

197. The fifth question took a similar approach to the "Teflon" style of question used successfully in the United States. However, Gault J found that the emphasis on "types" of cheese in the first question would have conditioned responses to subsequent questions by implicitly suggesting that "Vintage" was a type of cheese. Although the fourth question asked respondents about "brands" of cheese, Gault J considered that respondents would already have been predisposed to consider Vintage as a "type" of cheese, of which there were different "brands". Overall, His Honour concluded that, by this stage of the survey, respondents would have been primed to consider "Vintage" as a generic or descriptive word, thus their responses to the final question would also be likely to reflect the impression created by earlier questions.

198. Gault J's criticism is similar to the comments made in *Patience & Nicholson (NZ) Ltd v Cyclone Hardware Pty Ltd* (2001) 3 NZLR 490, where the repeated use of the words "*Patience & Nicholson"* was alleged to have increased the salience of this response.

However, while Hansen J did not accept this as a fatal flaw, Gault J was less sympathetic. Thus, although the questions used were technically not leading, the overall effect of the question sequence in *Mainland Products Ltd v Bonlac Foods (NZ) Ltd* (1998) 8 TCLR 224 was held to have undermined the survey. As a result, Gault J found it no longer comprised a fair test of whether the word "Vintage" was used by consumers to signify a brand or a product category.

199. Gault J raised similar points about the survey adduced in *Anheuser Busch v Budweiser Budvar National Corporation* (2003) 1 NZLR 479. He noted: *"A difficulty with the survey is that those who designed it do not seem to have been instructed on the points sought to be tested."* Related criticisms include the lack of a realistic purchase context, an issue of external validity discussed earlier.

200. Concerns about the survey focus and the extent to which this corresponds to the legal issue of interest were also raised in *Cookie Time Ltd v Griffins Foods Ltd* (2000, M1756/SW00) where confusion between tubs of cookies containing either Cookie Time products or Griffins' products was alleged. In this survey, respondents were not specifically asked about the two brands, but were instead questioned about how recently they had seen the product and the context in which they had seen it. Because the products were available through different distribution outlets (sold via personal selling in workplaces cf. through supermarket chains), and had been available for varying periods of time, responses that indicated respondents had seen the Griffins' product, more than three months ago, or that they had seen it offered for sale in their workplace, were interpreted as evidence of confusion. However, these results do not address the question of whether consumers would confuse the two brands (a question that was difficult to address because of flaws in the presentation of the stimulus material).

201. The need for a fair test has been apparent throughout this discussion and received detailed attention in *Klissers Farmhouse Bakeries Ltd v Harvest Bakeries Ltd* (1985) 2 NZLR 129, where Davis J considered consumers' likely confusion resulting from another brand using gingham styled marks on its packaging. His Honour was critical of the survey evidence adduced and indicated that this had not constituted a fair test of the issue to be determined. The survey design had involved showing respondents photographs of bread packages in which the brand names and other distinguishing characteristics had been hidden and so the gingham or checked pattern was clearly displayed to respondents. The lineup of photographs included one Quality Bakers brand among four Klissers' products.

202. The survey results showed that 46.8% of the people surveyed thought the checked pattern made the packages stand out and 75.5% thought the brands shown were all made by the same manufacturer. However, Davis J found the survey was not fair and reliable

because the stimulus material provided only one view of the packages and this emphasised the checked pattern at the expense of other product attributes. In addition, the photographs did not feature the total "get-up" and so did not allow a comparison of the brands on this basis. Finally, the survey did not explicitly test whether respondents would confuse the brands shown.

203. As a result of these limitations, Davis J concluded: *"My assessment of the whole of the evidence is that it is quite unsatisfactory as a basis for arriving at a conclusion on the issue of confusion based on get-up of the Klissers' packaging"* 156 [1985] 2 NZLR 143. In particular, His Honour found that the failure to present the entire packaging reduced respondents to a situation that did not parallel the situation they would normally encounter in store, where all the brands' attributes would be only display and available to inform their purchase decision. His Honour spent some time considering how bread might appear in display containers and whether other distinguishing features of the brands' packages would be available to consumers. Overall, on the basis of this analysis, he concluded there was not a reasonable likelihood that consumers would confuse the two brands.

204. This decision was subsequently appealed and more detailed discussion of the survey evidence was presented in the Court of Appeal judgment. Here the extent to which the survey met criteria established by Whitford J in *Imperial Tobacco v Philip Morris Ltd* [1984] RPC 293, was discussed in more detail. In particular, challenges were levelled at the nature of the questions used and the extent to which these represented a fair test of confusion, the artificiality of the survey context, which failed to replicate a normal purchase context, and the sample size. However, the Court of Appeal ruling held that failure to meet Whitford J's criteria did not necessarily render the survey evidence inadmissible in New Zealand. However, they did not issue a specific determination about the acceptability of the survey evidence, since they viewed this as a matter that should be determined when the evidence could be tested in cross-examination.

205. In the full judgment *Klissers Farmhouse Bakeries Ltd v Harvest Bakeries Ltd* [1988] 2 TCLR 555, the Court of Appeal upheld criticisms of the survey evidence, which they considered had *"been conducted in a way that tended to emphasise the checks and not other features inherent in the total get-up".* As a result, the findings from a test that had not been fairly conducted could not be relied upon to reach a just conclusion.

206. In addition to ensuring that the tests themselves are fair, researchers must also ensure that the estimates are fairly and properly reported. In IPONZ case 154636 (1994/05), survey evidence relating to the word "Ritz" was introduced. A university student conducted 119 interviews outside an Auckland shopping centre; she showed

respondents a copy of a page from Paton's Ritz Pattern Book on which Ritz patterns were shown and asked them what the "*Use of the name 'Ritz' on this product*" meant to them and how and when they came to hear about the name. Of the 119 people interviewed, eleven made an association between the Ritz hotel and the knitting patterns.

207. However, Assistant Commissioner McCardle noted that eight of the ten people who subsequently provided declarations did not make the association between the hotel and the knitting patterns in response to the first question, but when asked the second question. After examining the survey evidence, he noted that most of the initial responses to the first question comments on the word "Ritz" in an adjectival sense ('classy' or 'stylish'). He concluded that this information did not provide insights into whether *"purchasers of yarns and threads were aware of the opponent's use of the mark on such goods or would associate goods bearing the mark with the opponent"* (p. 7).

208. The same problem is apparent in many international cases where survey evidence has been adduced. The question of whether surveys can adequately represent the legal question of interest has also been challenged. *In Warner Brothers v. American Broadcasting Companies, Inc*. the judge rejected the evidence presented on the grounds that the survey questions were invalid because the general public were not able to apply to correct legal test to the issue. The judgment noted *"When a trial judge has correctly ruled that two works are not substantially similar as a matter of law, that conclusion is not to be altered by the availability of survey evidence indicating that some people applying some standard of their own were reminded by one work of the other*" (reported by Eko, 1998, p. 601).

209. In other cases, the likelihood that a consumer survey will assist the judge or hearing office is low, not because the survey is inherently unfair or incorrectly focussed, but because it was designed for a different purpose. Problems with the analyses led Wylie J to criticise the survey adduced in *Pitstop Exhaust Ltd v Alan Jones Pit Stop International Ltd* (1987) 2 TCLR 502. His Honour also noted that, even if the estimates adduced had possessed a clear logical validity, the survey itself was unhelpful because it had not been undertaken to assist the proceedings, but for some other purpose. This highlights the need to differentiate between evidence documenting brand awareness or market share on the one hand, and evidence testing confusion or distinctiveness, on the other. While the former can establish whether a mark has been known in a market, it does not provide direct insights into whether that mark is either distinctive or likely to create confusion with another mark. To investigate these latter issues, a different type of survey needs to be undertaken.

210. Similarly, in *Pioneer Hi-Bred Corn Company v Hy-Line Chicks Pty Ltd* [1975] 2 NZLR 422, Cooke J discussed questionnaires that had been sent to individuals associated with the poultry industry in New Zealand and concluded that *"All in all, the questionnaire procedure has not been notably successful in this case"* [1975] 2 NZLR 422, 430. The reason for his Honour's comments appears to be that many of the responses had not been verified by statutory declaration. Of those that were subsequently verified, the responses carried little weight since they examined the use of trademarks within New Zealand whereas the case involved marks used internationally. In addition, the questionnaires had been distributed some five and six years after the initial application and so could not offer insights into the meaning or use of the marks at that time. Finally, the sampling procedure had not been outlined.

211. These cases highlight the need to ensure survey questions are designed to inform the question judges must determine. Clarifying the legal points at issue in a dispute, and determining the questions that will get to the heart of this issue, requires a close working relationship between the survey researcher and the legal team. As the sections above show, researchers may use a variety of approaches in designing questionnaires. However, in addition to considering issues of fairness and balance, and ensuring the survey objectives match the question the judge or hearing officer must consider, researchers need also to consider consumers themselves, since the level of involvement consumers have in a decision may also influence the questions used in the survey. We examine the question of consumer involvement in the following section.

Consumer Involvement

212. Because the type of decision confronting consumers allegedly at risk of being deceived in passing-off cases may vary from case to case, researchers must also consider the level of attention consumers are likely to pay to the purchase decision at issue. This, in turn, may affect the context they create in their questions, the tasks they ask respondents to perform, and their interpretation of the legal issue.

213. The question of how consumers process brand logos, names and other elements of trade dress has been examined in several cases. Some experts have suggested that brands' effects are akin to gestalt, and that distinctive brand attributes are processed as shortcuts that simplify consumers' search and recognition processes (see IPONZ T30/2004). This type of processing is often referred to as "low involvement" and suggests that brand imagery may have a strong influence on consumers' choices, particularly if they do not expend great effort evaluating and comparing brand attributes before purchasing.

214. By contrast, other cases have made the point that consumers' decision making would be "high involvement" and would require more detailed analysis of the brands and the attributes these possessed. Decisions that involve deliberation and the collection and evaluation of evidence from a range of sources may be less likely to be affected by misleading claims or confusion over marks that are alleged to be deceptively similar.

215. This point was made by Cooke P in *ASB Bank Ltd v Trust Bank Auckland Ltd* (1989) 3 TCLR 77, where consumer survey evidence that attested to the level of involvement consumers were likely to have in decisions about bank accounts was adduced. His Honour summarised the arguments: *"the nature and commercial context of the service in dispute are such that there is no risk of the 'consuming' public being confused into thinking that they are getting ASB's HIT account when they open a TBA HIT account. The opening of an new bank account involves some deliberation and is not comparable with, for instance, casual or impulse or urgent purchases in a supermarket"* (1989) 3 TCLR 80.

216. Two surveys were adduced in *ASB Bank Ltd v Trust Bank Auckland Ltd* (1989) 3 TCLR 70; one examined staff members' assessments of whether consumers were confused by the two similar brand names and the second, a consumer survey, examined how consumers made the decision to bank at a specific financial institution. A telephone survey of 96 individuals holding HIT accounts found that none made the decision to bank with Trust Bank Auckland because of their HIT account. Furthermore, evidence that only two left Trust Bank Auckland believing it to be part of another bank was supplied; furthermore, these two respondents had not opened HIT accounts (1989) 3 TCLR 81.

217. Although both surveys were accepted, more weight was placed on the consumer survey. However, Cooke P commented that the survey did not examine whether respondents perceived a relationship between the ASB HIT account and the Trust Bank Auckland HIT account. That is, the survey did not appear to address the question of whether consumers mistakenly confused the two accounts or believed them to be offered by the same institution. This case highlights the need to integrate survey design factors; in this instance the survey must both recognise the type of decision and the context in which this occurs, while still ensuring that the legal issue is considered.

218. For example, in *Automobile Club de L'Ouest, Aco v South Pacific Tyres New Zealand Limited* CIV 2005 485 248, the expert for Dunlop suggested that consumers could have more than one level of association with the words "Le Mans" and indicated that it was important to differentiate between primary and secondary associations.

219. Similarly, in *Universal College of Learning v ACP Computer Solutions Limited and The College of Future Learning New Zealand Limited,* CP 9-01, Oct 22, 2003, Hammond J

commented that the decision to select a tertiary institution at which to study was *"not at all like buying cereal off the shelf – the decision is unlikely to be based simply on a name or just the look of a brand"* (para 47).

220. This section has examined structural factors researchers need to consider in question wording and questionnaire design. These factors all influence the perceived fairness, and hence validity and accuracy, of the survey estimates. The challenges levelled at survey evidence also highlight the importance of using quality assurance procedures to reduce the likelihood of serious measurement error. One such approach involves controls, which provide benchmarks that estimates of distinctiveness or confusion can be measured against.

Controls

221. Controls involve the use of independent examples or holdout groups that are either not related to or not exposed to the disputed mark or claim. They may assist in identifying the level of guessing that has occurred, particularly where the control represents a plausible response that is not available in the market. For example, use of a control brand that has not been involved in any alleged deception, or a control colour that is not associated with a particular product category, may help identify the level of guessing that has occurred and enable researchers to present two sets of estimates: those based on the full sample, and those that have been discounted by the level of response to the control items.

222. Controls may assist researchers to account for other types of confusion that consumers might experience. This has been particularly relevant in US cases involving disputes over domain names. Where a part of the name in dispute is used by other traders not involved in the action, the plaintiff must demonstrate that the defendant's behaviour creates confusion above and beyond the noise that might ordinarily be expected in the product category. Controls thus assist researchers to establish benchmark levels of noise or confusion; from there, they can assess the "noise" or background confusion that exists within a market and can explore the extent to which this is exceeded by the allegedly confusing behaviour.

223. However, not all researchers agree on the value of control measures. Liefeld (2003), for example, criticised the use of fictitious brands or attributes and argued that consumers could not respond thoughtfully to questions that asked their opinion on brands that do not exist. However, despite Liefeld's criticisms, consumers' willingness to make responses or attribute associations may nevertheless provide a check on the extent to which their responses are based on actual knowledge or experience, and may support the

use of "don't know" options. Moreover, the courts have sometimes criticised surveys for failing to include proper controls.

224. For example, in a US case, *Simon Property Group LP v. mySimon*, Inc 104 F. Supp. 2d 1033 (S.D. Ind. 2000) which involved a dispute over a domain name, the court criticised the survey proposal for failing to include controls that could assess whether other marks that included the word "Simon" confused consumers. Because of the lack of controls, the court concluded that the proposed evidence would not offer insights into whether the confusion resulted from similarities in the services offered by the two parties, or from similarities in the name.

225. Similar issues have arisen in New Zealand cases. For example, in *Magellan Corporation Ltd v Magellan Group Ltd* (1995) 6 TCLR 598, a dispute arose over use of the name "Magellan" by two companies. Magellan Corporation sent circulars to 400 people and organisations in an attempt to make clear the differences between the two companies. Subsequently, Magellan Corporation interviewed 101 people from the 400 sent circulars and found that the letter was recalled by only 69% of the sample; in addition, respondents' knowledge of the two companies was examined and the findings suggested some confusion over the company names, owners and locations.

226. However, Fisher J noted that he approached the survey findings with "caution" as "*No attempt was made to see what confusion there would have been between industry members which did not have similar names"* (1995) 6 TCLR 598, 610. This suggests that the survey's failure to include a control company name, which could serve as a benchmark for establishing confusion, undermined the credibility of the survey results.

227. Yet, conversely, in IPONZ case T43/2003, the survey adduced by Effem Foods Limited in support of their application to register the colour orange in relation to rice and rice snacks was criticised because a control colour had not been included in the experimental design. However, the hearing officer did not make note of this criticism in his decision and described the survey evidence as: "*proper, reliable, logically probative, and given in accordance with the rules of natural justice*".

228. Although the comments made about the Effem survey suggest that control questions or responses were not considered vital, use of controls can support survey evidence. For example, in IPONZ case T32/2002, which examined residual awareness, a telephone survey of 1000 people aged 18 years and over was conducted to test latent awareness of the Felix cat brand among members of the general public. The survey used a split-sample approach where half the respondents were asked questions about the word "Felix" while

the remaining 500 were asked about the word "Sylvester", which was employed as a control.

229. Although Assistant Commissioner Howie did not comment in detail on the survey design, he accepted the survey, which he noted had "*been presented in a professional and competent manner*" (p. 10). Moreover, the use of a control supported claims that the word "Felix" had a residual association with cat food, since 20% of those surveyed associated this word primarily with cat food (and 44% in total associated Felix with cat food) while less than 1% (2 from the 500 surveyed) associated Sylvester with cat food.

230. This decision was subsequently overturned in *Friskies Ltd v Heinz-Watties Ltd* 2 NZLR 663, although the survey evidence was not dispute. Following the success of Dalgety Spillar Foods Limited's application to remove Felix from the trademarks register, Dalgety subsequently applied to register Felix as a mark belonging to them (IPONZ T28/2004). Heinz Wattie objected to this application, and the IPONZ hearing saw the same survey evidence relating to residual awareness of Felix introduced.

231. Assistant Commissioner Walden accepted the survey evidence adduced in the earlier hearing, noting that the sample comprised a cross-section of the relevant public and that the interviewer instructions had been properly disclosed. He found that there would be an *"almost…inevitable risk of deception and confusion to the general public"* if Dalgety's application to register Felix were successful. The residual level of awareness was held to be sufficient to be likely to lead to unacceptable levels of confusion among the public, despite the fact that the mark Felix had been removed from the trademarks register.

232. Overall, neither closed questions nor open-ended questions can eliminate the possibility that some respondents will offer guesses in order to provide a response; nor will including a control option necessarily satisfy the courts that researchers can identify the true level of distinctiveness or confusion. However, careful design of the question, inclusion of a statement that legitimises a "don't know" response, and use of controls, nevertheless provide safeguards that will go some way towards deflecting criticisms that the survey estimates lack validity. Another key quality assurance mechanism aims to reduce error that poorly trained interviewers can introduce.

Interviewer error

233. A primary source of measurement error is through interviewers themselves, particularly if they fail to follow instructions, become over-zealous in their pursuit of a particular response, or fail to demonstrate the level of professionalism necessary. For example, if interviewers do not administer the questionnaire as instructed, or if they

imply through their tone or through other gestures that some responses are preferable to others, they may influence respondents' behaviour.

234. Morgan (1990) suggested that interviewers should know neither the survey client nor the purpose of the research, and the questionnaire design should minimise or eliminate the need for interviewer intervention in the survey (see also Preston, 1992). While it is difficult to withhold information about the survey topic, since this can usually be discerned from the questions themselves, interviewers should not know the party that commissioned the research. There are specific cases where interviewers have been fully apprised of the survey client and, in a mis-guided attempt to do a "good job" for the client, have engaged in a level of probing that goes beyond normal practice, apparently in an attempt to elicit "helpful" responses.

235. Some researchers have even stressed the desirability of a triple blind interview situation where the fieldwork supervisor is not aware of the client, and cannot disclose this to the interviewers, who in turn are unable to convey this information to respondents. This level of non-disclosure would ensure the researchers were able to defend allegations that knowledge of the survey client had influenced interviewers' behaviour in any way.

236. Interestingly, in New Zealand, surveys conducted by staff employed by the attorney acting for one party in a dispute have been accepted, despite the obvious knowledge those staff would have of the proceedings. For example, the ROBOCUP case (IPONZ T51/2002), the Philips electronics case (IPONZ T29/2003),and the "Le Mans" case all involved small surveys conducted by staff of the law firm acting in the application. In the former two cases, the survey evidence was not critical to the outcome, thus although both surveys were admitted, other limitations meant the quality of the interviewing was not considered in detail. Similarly, the use of questions that "*were not entirely open"* in the "Le Mans" survey meant more technical factors, such as the interviewing, did not play a definitive role in the decision.

237. Glazebrook J made more detailed comments about the quality of the interviewing in *Cookie Time Ltd v Griffins Foods Ltd* (2000, M1756/SW00). The survey adduced in this case was conducted by a consultant to the plaintiff and the plaintiff's marketing manager. Although Glazebrook J accepted the survey, he noted that the survey was "*an exercise that had been undertaken in haste*" that had "*shortcomings*" and that "*needed to be heavily discounted*". Nevertheless, he found the survey provided a "*possible indication of a level of confusion*" but suggested it would not carry weight at the full trial unless "*redone or at least validated by market survey professionals*" (para 48).

238. Similar criticisms were levelled at the evidence of interviewer training in *Commerce Commission v Griffins Foods Ltd* (1997). In this case, the field force supervisor had provided information about the briefing given to Auckland interviewers, but information about the training provided to interviewers from the other five centres where data were collected was not provided. However, Boshier J accepted that the research methodology was reasonable and did not comment specifically on the quality of the data collected and analysed.

239. Concerns over the quality of fieldwork undertaken highlight problems such as declining response rates that survey researchers and the interviewers face. Issues such as low response rates may lead interviewers to adopt short-cuts that increase the number of completed interviews they secure. It is vitally important that these short-cuts, which may see the abandonment of specified recruitment procedures, truncation of some question wording, or assumptions about respondents' likely answers, are never practised during the data collection.

240. The increasing difficulty of securing successful interviews and the need to ensure that the data are collected properly suggest that the prudent course of action is to use independent professional Interviewers. However, while interviewers must receive detailed instructions about the survey, they should not be advised who the survey client is or how the results will be used, since this knowledge may influence the behaviour of even the best trained interviewer.

241. Cases where the interviewers are aware of the survey purpose have been criticised both internationally and in New Zealand. For example, in the UK Black & Decker case it became apparent that the interviewers were not well trained and had been specifically instructed about the purpose of the survey. More critically, they knew that Black and Decker considered obtaining trademark protection for the colours they sought to protect as extremely important to their corporate strategy. This knowledge severely compromised interviewers' ability to conduct the interviews in a disinterested manner.

242. Similarly, the interviewers in *Anheuser Busch v Budejovicky Budvar National Corporation* [2001] 3 NZLR 666 were aware of the survey purpose and the tape recorded interviews revealed some intense probing as interviewers explored the associations respondents made between the two brands. However, Doogue J saw these issues as technical rather than substantive, and his overall concern was whether the survey fairly assessed the legal question of interest. His view that the survey evidence was not valid rendered the technical criticisms less important.

243. To reduce the impact of interviewer error, careful training should be undertaken. Ideally, interviewers would not have to make judgments about any aspect of the interview. Their training should therefore include details of how they must approach, recruit and select respondents. The questionnaire should include information about the sequence of questions they are to follow and any skip instructions that may require them to follow a different route through the questionnaire.

244. Details of the training provided should be recorded on video so that the judge or hearing officer can see the nature and extent of the training and each interviewer's aptitude for the task at hand. In *Levi Strauss Co. v Kimbyr Investments Ltd* (1994) 1 NZLR 332, Williams J commented on the *"particular care"* paid to interviewer training, which included provision of a video tape prepared by the project director that was sent to all locations to form part of the interviewer training.

245. However, despite the detailed interviewer training provided, Kimbyr criticised the survey adduced by Levi Strauss because a number of the questionnaires had been incorrectly completed, or were incomplete. Counsel for Kimbyr argued that, because errors in the questionnaire had been noted, the questions used must have been flawed, and that if the questions were flawed, the survey was also unreliable. Williams J ultimately did not accept these arguments because of two factors: evidence that the errors in the questionnaire were not such that they would have had a material effect on the estimates reported, and expert evidence that the questionnaire had been soundly constructed. Furthermore, Kimbyr's failure to call an expert who could provide a contrary view on the survey's validity and reliability was seen as a limitation to their case.

246. In addition to documenting the training provided by the project director, it could also be useful to consider providing evidence that field supervisors have role-played the interview with each interviewer and provided critical and constructive feedback on the standard of interviewing. Evidence of each interviewer administering a perfect interview to the field supervisor could also help establish the overall quality of the interviews undertaken, and the ability of the interviewers to follow the instructions and question sequence.

247. The instructions provided should make clear that interviewers must follow the questionnaire script and that they are under no circumstances to improvise, offer comment on respondents' answers, or suggest response options that are not included in the questionnaire. The interviewer instructions should clearly document a procedure used to recruit respondents. Ideally, this would be a randomised procedure where

interviewers must approach every nth passer-by, call in at every nth household, or where the telephone numbers they call are randomly generated by a computer programme.

248. Field force supervisors should be present throughout at least part of the interviewing so they can testify that the recruitment procedures were properly employed. Where mall-intercepts are undertaken, casual period observation by the supervisor, project director or consulting expert could also provide assurance about the quality of the data collection processes.

249. Researchers following the procedures suggested above should document each step of the survey process and ensure the training is clearly documented. Development of clear written documents outlining the procedures used will be important if the interviewers are required to submit affidavits and present themselves for cross-examination. The status of evidence provided by interviewers was discussed in detail in *Noel Leeming Television Ltd v Noel's Appliance Centre Ltd* (1985) 1 TCLR 283. Holland J was provided with evidence from 6 of the 7 interviewers who administered a survey relating to this case. Counsel for Noel's Appliance Centre Ltd challenged the evidence and argued that it was hearsay. However, Holland J found that the interviewers' evidence provided the basis for comments subsequently made by the project director, who was acting as a survey research expert. His Honour noted: *"The important evidence was the opinions and conclusions of those specially trained in market research based on these questions and answers. It is those conclusions and opinions which are of assistance to the Court"* (1985) 1 TCLR 286.

250. Measurement error can critically undermine survey evidence and can occur at many different stages of the survey design and administration process. We list some guidelines designed to assist researchers to prepare fair and balanced questions that are legally probative, administered to the correctly designed and selected sample by well-trained and competent interviewers. The following section overviews some of the academic literature that discusses approaches researchers have used to establish secondary meaning and confusion levels.

5. Establishing Secondary Meaning

251. To establish secondary meaning, researchers must demonstrate that a mark or name has become so distinctively associated with a brand that it is able to function independently to identify the brand. Once a mark becomes so strongly linked to a brand, its use by other traders would be likely to mislead or deceive consumers. Cases where survey evidence is adduced to support a trademark application thus need to establish that the attribute for which registration is sought is able to function as a source in its own right. This section begins by examining the concept of genericism and considering the research approaches that have been taken to estimate a mark's capacity to distinguish a particular brand.

5.1 Genericism and Distinctive Marks

252. When considering whether a mark is generic and should be available to all traders, or distinctive, and thus aligned only with one brand, judges and hearing officers review several factors. McCarthy (1998) suggested they should consider four pieces of evidence that will assist them to determine whether a mark functions as a "badge of origin":

- The presence or absence of the term in dictionaries and its definition in relationship to the goods;
- The manner of use by the trademark owner, i.e., whether the term has been used as a trademark or a common name for the product in question;
- The way in which competitors and the trade describe the product and the words they use to refer to it;
- Consumer surveys (sourced from Taylor and Walsh, 2002).

253. Simonson (1994) also suggested that the courts consider several factors when determining the primary significance of a name, although his criteria were based less on legal issues than on consumer perceptions. They included the share of buyers using the term as either a general description or a trademark; the proportion of buyers who use the term as both a general term and a trademark, and the dominant use of the term by this group.

254. In addition to reviewing both the word and how it is used, the courts also consider the availability and knowledge of alternative names, and the efforts made by the trademark owner to prevent the term from becoming generic (Taylor and Walsh, 2002, p. 102). These criteria make it clear that consumers' impressions are important considerations, although the competitive environment in which the mark is used also plays a role in establishing its status.

255. Genericism refers to the process whereby a mark is used to describe a product category rather than a specific brand. Trademarked names can become generic as a result of consumers' natural tendency to adopt words that convey in a simple and straightforward manner the product they wish to request or talk about (Swann, 1980). When considering whether a word is generic or eligible to be registered as a trademark, Swann noted that the courts have considered competitors' rights to use these words when they become common parlance, and the manufacturer's right to maintain exclusivity over a word that still contains what he describes as *"source identifying significance"* (p. 358). He argues that the appropriate test of genericism is not consumers' use of a word, but the extent to which they understand the word's use in a trademark sense.

256. The distinction between genericism and secondary meaning was first tested in the US when Bayer Company's registration of the mark "Aspirin" was challenged. Researchers assessed consumers' understanding of the word "aspirin". If they considered this denoted a product made by Bayer, then the name should remain under trademark protection. However, if they considered the term denoted a product that offered pain relief, then the word had clearly come to denote a product type, rather than a brand source.

257. Although those in the trade recognised Bayer Company as the source of Aspirin and had an alternative technical description available to describe competing products, the court noted that consumers did not have well-known or easily recognised alternative descriptions available to them. Thus, despite the trade's ability to use alternative descriptions, the court ruled that Aspirin should lose trademark protection because maintaining this status would inhibit competitors' rights to promote their products in a manner that consumers would readily understand.

258. Among other things, this case highlighted the importance of considering relevant audiences and the differing levels of knowledge they may possess. Thus, while consumer survey evidence may be persuasive in cases involving genericness, surveys of the trade have carried less weight if the trade are not considered to be the relevant public, and are merely offering hearsay opinions about end-consumers' views.

Question Wording

259. In reviewing consumer survey evidence, US courts have also expressed strong reservations about the question wording used. For example, questions that require respondents to make a "*false dichotomy*" between brand and generic names have not been accepted as the courts have regarded these questions as susceptible to variations in question wording (Taylor and Walsh, 2002). The courts have also noted the

importance of providing respondents with definitions that clarify the difference between product and brand names. Failure to do this has led them to conclude that respondents are not offering an opinion on the meaning of the word, but are instead being asked to draw a legal conclusion (Taylor and Walsh, 2002, pp. 103-104).

260. Folsom and Teply (1988) clarified the distinction between questions that explore whether a word or mark functions as a trademark or as a generic term. They noted that a generic term answers the question "what is the name of this product?" while a functioning trademark answers the question "who is the source of this product?" (p. 4). Swann and Palladino (1988) proposed a similar distinction when they wrote: "*A brand name generates particular expectations about a product and distinguishes it from the products of other companies. A brand name is an assurance that a product will consistently possess certain quality and other characteristics. A generic name does not generate particular expectations other than as to basic product type. A generic name equally describes all products of the same basic type*." (p. 193).

261. This distinction has led to the development of surveys posing "product-category" questions that explore how consumers describe or request specific types of products, and "brand awareness" questions that ask respondents to classify particular words as either brand or generic names, or to identify the brand names associated with a particular type of product.

262. However, as Folsom and Teply (1988) noted, these surveys assume consumers either use a term generically, or use it as a trademark, to denote the source of a product. As they point out, consumers may use some terms in both senses. They write: *"Such consumers thus use and understand the word as a hybrid in everyday discourse to refer to a product class, yet are aware of its brand significance"* (p. 7).

263. To date, the main approaches used to test whether a mark has acquired secondary meaning have assumed the mark functions either as a trademark or as a generic term. We consider the key approaches used in the United States, where consumer survey evidence has been adduced over a longer period of time, in the following section.

5.2 Approaches to Estimating Secondary Meaning

264. US researchers have developed two main techniques to assess whether a mark has acquired secondary meaning. The first of these is loosely known as the purchase encounter approach, which was used *in American Thermos Products Co, v Aladdin Industries, Inc* 207 F.Supp. 9, 20-22 (D.Conn.1962). Respondents administered this approach are asked a range of questions that first examine their knowledge of the

product under investigation and then explore the type of store at which they would purchase such a product. They are next asked if they were to enter such a store, what they would ask an assistant for, should they wish to purchase the product. They may also be asked if they can describe other ways in which they might ask for the same product. Where a large proportion of respondents use the mark or term in their answer, the results are interpreted as supporting an argument that the term has become generic. By contrast, if the term is not frequently used in responses, it is assumed to apply to a specific brand rather than to the product category as a whole.

265. The Thermos case paralleled earlier disputes where consumers have used well-known and dominant brands to signify the product category, although Thermos had arguably taken more care to use the name "Thermos" in conjunction with the words "vacuum bottle". However, despite the American Thermos Products Company's efforts to promote the trademark status of "Thermos" the courts ruled their efforts were too late and largely ineffectual, since Thermos had already become a household name. In issuing their decision, the court noted that their primary concern was to protect consumers from deception, and that only once they had satisfied this requirement could they turn to examine the implications on traders of continuing with (or removing) monopoly rights to a trademark.

266. However, Swann (1980) criticised the survey evidence used in Thermos and argued it failed to assess the principal significance of the term and relied unduly on how the term was used. He pointed to the conditioning effect early usage questions would have had on respondents' answers to subsequent questions and suggested that question order effects are highly likely to have occurred. Thus, while the individual questions were not considered to be leading, the overall effect of the question sequence allegedly promoted a conclusion of genericism (or loss of trademark significance).

267. Simonson (1994) reached a similar conclusion, for the reasons Folson & Teply (1988) had outlined. Like them, he found the Thermos approach favoured genericism as the questions imply respondents use a term either as a trademark or as a generic term, when in fact they may understand both applications of the term, and use both correctly. Concerns that the "Thermos" questions could be inherently biased, led to the development of alternative methodologies. Of these, the "Teflon" approach has received detailed attention in the literature.

268. The "Teflon" case, used in *E.I.du Pont de Nemours Co., Inc. v Yoshida International Inc.* 393 F.Supp. 502, 185 USPQ 597, 603 (EDNY 1975), involved two surveys. The first of these followed the Thermos approach and found that of women who were aware of non-stick coating, 80-85% responded with "Teflon" when asked to name the product; 60% -

70% indicated that "Teflon" was the only name they would use to describe the product, and less than 10% identified DuPont as the manufacturer of Teflon. However, the courts found that the questions used sought a name from respondents without clarifying whether respondents understood this name to imply a type of product or the source of the products. The second survey provided respondents with a definition of brand and generic (or common) names, and then asked them to classify different terms according to the definition they had been given. The court found that this approach enabled an assessment of the principal significance of "Teflon".

269. Swann (1980) suggested that the Teflon Survey addressed many of the shortcomings he identified in the Thermos survey, since it did not favour genericism. However, the very process of providing respondents with definitions and asking them to classify terms according to these definitions meant their responses may have reflected their ability to use those definitions, rather than their actual usage of the terms in dispute. Folsom and Teply (1988) made a similar point when they noted that the different questions used in the Thermos and Teflon cases appeared to affect the results obtained. Whereas 75% of those in the Thermos survey described the product as a "thermos" only 46% of those in the Teflon survey described it as a common name. In addition, only 12% of those in the Thermos survey indicated that Thermos was a trademark, whereas 51% of those in the Teflon survey called it a brand name (p. 7, note 13).

270. These differences are difficult to explain in terms of sampling error and almost certainly reflect differences in questions used. Simonson (1994) suggested that, while the Teflon survey had eliminated the bias towards genericism that undermined the Thermos approach, it may have replaced this with a bias that favoured a distinctive outcome. Overall, neither the Thermos nor the Teflon approaches appear to provide unbiased estimates of the level of distinctiveness a mark has. More importantly, they may not address the legal question of interest.

271. In their review of the Thermos approach, Swann and Palladino (1988) noted the limitations of questions that ask consumers how they would describe or ask for a particular product and pointed out that these questions do not offer any insights into how consumers actually use the disputed terms. Furthermore, they argued that the questions themselves may prompt respondents to provide a brand name as an answer, without first having considered whether this name is the primary way in which the word functions. If their reasoning is correct, it would imply that the questions are not providing insights into the legal issue judges must address.

272. Swann and Palladino also evaluated the questions used in the "Teflon" survey and discussed the criticism that the questions used may lead respondents to classify unusual

sounding names as brand names. They concluded that if terms were ambiguous, respondents would use a "Don't know" option to indicate that they could not answer the question. However, they considered that asking respondents to differentiate between brand, generic and hybrid names would be confusing, and argued that respondents would not understand this distinction easily. They noted: "*The differences among brand, generic and hybrid terms may be particularly elusive to survey interviewers and respondents*" (p. 183).

273. Leiser & Schwartz (1983) also recognised concerns about the style of questions used in the Teflon survey and offered a series of suggestions to improve the robustness of these. First, they suggested that researchers make it clear that products may have more than one generic name, and that the fact a product is made by only one company does not mean it is necessarily a trademark. They also recommended use of a list of terms, as used in the original Teflon survey, since this would reduce the likelihood of criticism that respondents have been asked to answer a question they do not fully understand. In addition, they suggested this format provides a control for what they describe as the "noise factor"; essentially, this means that the use of other names enables researchers to establish a benchmark against which results for the test term can be compared.

274. Leiser & Schwartz (1983) also outlined a third approach, which they described as the motivation approach, that could replace or supplement the Thermos and Teflon approaches. The motivation approach examined why respondents had purchased a particular product and explored whether their purchase arose because they liked the manufacturer's products, or because they liked the product, irrespective of who produced it. If consumers are motivated to purchase because of a particular manufacturer, the mark would be seen as indicating a source. However, if consumers simply liked a product in general, the mark would be more generic, since the actual source of this would not be the primary motivating factor.

275. Perhaps predictably, this approach has been criticised for failing to address the legal question of interest, which is how respondents regard the term – as a generic or brand name – rather than what motivates their purchase. (Leiser & Schwartz (1983), p. 388). Despite this, Leiser & Schwartz (1983) suggested marks that have a strongly motivating effect on consumers' behaviour are likely to be those liable to infringement and thus worthy of trademark protection.

276. Although Leiser & Schwartz (1983) supported the use of control questions, Swann and Palladino (1988) criticise the proposed use of these. They suggested additional questions that appear to explore the same issues may irritate or confuse respondents, and are thus

not likely to provide an appropriate benchmark against which response distributions to earlier questions can be assessed.

277. Swann and Palladino (1988) summarised the range of marks available when they described these suggest as existing along a continuum, anchored at one end by novel words that refer to a specific brand and that are rarely if ever used in a generic sense. At the other end of the continuum are words that although once used to denote a particular brand have now become so widely used to describe a product type that they have become fully generic. Between these two points, Swann and Palladino suggest more ambiguous terms, that can be used as both trademarks and generic product category descriptions exist.

278. To Swann and Palladino (1988), confusion of these two categories of respondent, which exhibit quite different understandings of a term's trademark significance, is a fundamental flaw that researchers have yet to address in full. Their claims that that generic responses may arise from shorthand use, lack of attention to the question from respondents, and what they describe as the "*supply a name*" syndrome (p. 189), where respondents supply any answer rather than a "*don't know"* response, suggest further work into respondents' interpretation of questions is required.

279. The on-going debate between Folsom and Teply (1988) and Swann and Palladino (1988), centres on the purpose of the questions used and the extent to which these assess use rather than understanding. Swann and Palladino (1988) noted that use may reflect laziness and expedience, rather than knowledge of a term's actual status. Because of this they maintain that survey reviewers need to consider whether the questions used are inherently biased to arrive at a generic response.

280. However, distinguishing between *"a casual generic usage of a term by a respondent who has a clear understanding of the term's brand significance and a fundamentally generic usage of the term by a respondent who has only a hazy comprehension of the term's trademark function*" (p. 188) remains a challenge. While a number of research methodologies exist and have been successfully adduced, they have not achieved widespread endorsement and concerns that they may contain systemic biases remain.

281. In addition, the question of what threshold should be reached before secondary meaning is established must also be addressed. Folsom and Teply (1988) suggested a 50% threshold should be used to determine whether a disputed word is generic or entitled to trademark protection. Following this initial assessment, they recommended examining the extent to which a word is used in a hybrid sense, as both a trademark and to denote a generic product type. They also noted the importance of ascertaining whether

alternative words exist, should a formerly generic word have come to denote a specific product source. This latter assessment examines whether granting exclusive rights to a word will impede the competitive dynamic of a marketplace.

282. In New Zealand, the question of what level of distinctiveness needs to be established for a trademark application to succeed remains unclear, although the lower the level of inherent distinctiveness, the higher the level of association required. For example, applications to register specific colours have failed where the level of brand-colour association has fallen below 50% (T30/2003, Frucor application to register the colour green; CT43/2003, Effem application to register the colour orange). In the former case, even though survey evidence suggested that 65% of those interviewed who had used the product category in question (energy drinks) associated lime green with the brand "V" (a total of 59% of the sample of 500 declared they had purchased from the energy drinks product category), the application to register the colour green for this category was not successful. Assistant Commissioner Brown QC noted that *"Given such a low level of inherent capability of distinguishing, it is my view that a significantly greater level of factual distinctiveness is required in this case to justify registration of the trade mark than has been demonstrated…".*

283. Cases involving secondary meaning need to establish that consumers use the mark in question to identify a single source for a product. McCarthy suggested that, if 50% of those surveyed made the association, this should be sufficient to establish a case for the mark to be registered as a trademark. He also suggested that as well as ascertaining which source respondents associate a product with, that researchers also examine the reasons why respondents associate the product with that source. However, investigating the reasons why consumers make associations or hold perceptions is difficult as there is no way of validating the responses they provide. Nevertheless, judges may make inferences about consumers' reasoning, as occurred in *Klissers Farmhouse Bakeries Ltd v Harvest Bakeries Ltd* (1985) 2 NZLR 129.

284. Distinctiveness requires not only that a considerable proportion of those surveyed associate the attribute with a particular brand or brand stable, but that the attribute (or mark) is somehow peculiar to the brand or brands in question. This issue arose in *Klissers Farmhouse Bakeries Ltd v Harvest Bakeries Ltd* (1985) 2 NZLR 129, where ownership of a gingham pattern used in a particular format on bread wrapping was in dispute. Davison J found that, although survey evidence reported that 40.9% of those interviewed considered that Klissers' brands were distinctive because of the gingham look, only around a third this proportion (15%) considered they particular use of the pattern at the top and bottom of the wrapper as distinctive. As a result, His Honour concluded that: *"the identification of Klissers' packaging …has been mainly on the basis that it is the*

only one known by many people as having gingham checks on its packaging and not by reason of a distinctive get-up of the packages" 153 [1985] 2 NZLR 143.

285. These decisions recognise the need to design surveys that not only test whether distinctiveness has been established, or whether confusion would occur, but that do so in a fair and robust manner. In particular, the Klissers discussion reveals the need to display packaging so that consumers encounter a situation similar to that they would find in a retail outlet, rather than one that has been engineered to give greater prominence to some aspects of the packaging.

286. In T63/2002, AMI Insurance objected to an application to register the name "Kelly Brown" and "Kelly Brown Beer" on the grounds that this would cause confusion with an advertising image they had created and promoted heavily, and that that they argued was widely associated with their company. To oppose the application, AMI Insurance needed to demonstrate that a substantial number of people were aware of the mark, and thus could be at risk of being confused and deceived should the mark be used by another trader.

287. Although sales figures and details of advertising and promotional support are often used to establish the extent of a company's reputation, AMI also adduced a survey of consumers. This survey was designed to test the level of association respondents had with this name and concluded that registration and use of the name by a beer company would create confusion. However, although Assistant Commissioner Frankel, described the survey as "*one of the better ones*" (p.22) she had seen in these types of proceedings, she paid little attention to it, except to note that the survey findings indicated respondents made a "*significant connection*" between a beer and soft drink named "Kelly Brown" and AMI, a factor that appeared to confirm her decision to decline the application to register "Kelly Brown" in connection with beer.

288. These cases suggest that the threshold level of distinctiveness depends very much on the mark in question and the overall structure of the market. Where the mark is highly distinctive and the brand seeking registration is well-established and has high penetration, the evidence of distinctiveness would seem comparatively easy to establish. By contrast, where the attribute has an inherently low level of distinctiveness, and the brand has lower penetration and does not dominate the product category, the level of association would need to be higher to justify trademark registration.

289. In addition to estimating distinctiveness and the proportion of consumers or likely consumers associating a mark with a particular source, researchers have also conducted surveys to estimate confusion, which may occur when trademarks are mis-appropriated or

deceptively similar marks are created and used. The following section examines this research question, and the approaches that have been used to address it, in more detail.

5.3 Estimating Confusion

290. Confusion occurs when consumers attribute meanings to claims that are not literally correct, or when they regard a mark as denoting a source that is not the true source of the brand bearing that mark. Survey evidence is useful in these cases as it enables evidence of consumers' perceptions and behaviours to be brought to the court. Pflüger (2001) summarised this view when she noted the importance of survey evidence as a means of offering insights into a market that a judge or hearing office may not be able to provide, even if they belong to the group of consumers at risk of being confused. As Pflüger noted, judges' experience in hearing cases means they bring a more experienced and knowledgeable perspective to bear on cases involving consumer confusion. Ironically, this specialist knowledge may prevent them from seeing a situation in the way that ordinary consumers would see it.

291. Consideration of whether all consumers merit protection or whether some groups should be regarded as more vulnerable than others has received detailed attention. Pflüger (2001) noted the different standards used to determine the type of consumer researchers should examine. She comments on German law, which initially examined consumers who were "*casual and inattentive*", but that now considers the "*average informed consumer*" (p. 3). This move from consumers who may be paying little attention, and who may therefore be at greater risk of confusion, to a group who is expected to have reasonable knowledge of the product category in question, reduces the level of protection afforded. However, as Pflüger noted, this change raised new questions, including the definition of "*informed consumers*" and the basis on which decisions regarding confusion will be made.

292. Likelihood of confusion surveys assess whether consumers are likely to confuse two brands, one of which is allegedly deceptively similar to the other. These surveys typically explore the source of the brands before examining the attributes that led respondents to associate each brand with a specific source. In practice, these surveys begin with a question such as "*who puts out this product*?" and follows this with a probing question: "*What makes you say this*?" Other options include placing the disputed products in a line-up with a control brand and asking respondents whether the products are made by the same or different companies. Again, a probing question is used to explore the reasons why respondents offered a specific answer.

293. Simonson (1994) identified several methods for estimating the likelihood of confusion between different trademarks. He outlined the "top of mind" approach, which involves

showing respondents the junior (or allegedly confusing) mark and asking them what it mark brings to mind for them. If respondents reply with the name of a company, this is recorded. If not, they are asked a specific question that explores any associations they might make with a particular company. This approach also asks respondents to explain the association they made so researchers can understand the link between the mark and company (where such a link was made).

294. Where the senior mark is mentioned as part of the explanation, these responses are used to calculate the likelihood of confusion that exists. However, as Simonson pointed out, this approach has several limitations, particularly the fact that respondents' recall may have little to do with whether they confuse the two marks. Second, Simonson noted out that this method is vulnerable to question order effects, where the presentation of questions put to respondents may shape their response to subsequent questions. That is, if respondents are uncertain of a specific response, they may guess and the answer they provide is likely to have been shaped by the content of earlier questions.

295. In addition, respondents' own experience with the product category can lead them to nominate brands they are familiar with, especially when they have little experience of the brands at the heart of the dispute. Those respondents with little direct experience of the product category are therefore likely to rely on the context created by earlier questions or on their overall perceptions of the product category. In this case, large brands are likely to be disproportionately offered as responses, in line with Double Jeopardy effects.

296. Simonson (1994) also described a second approach, known as the company identification approach. This method involves showing respondents a junior product and asking them three questions: who they think puts out the product, why they offered the response they provided to the first question, and what other products they believe the company they nominated in the first question manufactures. As Simonson noted, this method examines the question of confusion directly and does not rely word-associations as a proxy measure of confusion.

297. However, Simonson pointed out that this method is most likely to reveal confusion if respondents fail to distinguish between the two marks, but are able to identify the source of the senior mark. As a result, he pointed out that respondents are less likely to exhibit confusion if they do not know the company that manufactures the senior mark. He also suggested that this approach may discourage respondents from offering "don't know" responses and increase the probability that they will offer guesses in which they have little confidence. Simonson also questioned the external validity of this approach and argued that, in an actual purchase situation, respondents may well have made more

effort to clarify the relationship between the two marks. Surveys that fail to replicate the actual buying circumstances that consumers are likely to encounter have often fared poorly in the courts.

298. A third technique Simonson analysed is the company identification forced-choice approach, where respondents are shown both marks and asked whether they think the same or different companies put these out. However, as with the company identification technique, Simonson noted that this method may also underestimate confusion if the names are confusingly similar and if it would be illogical for a single company to use both names.

299. Furthermore, presenting respondents with both marks and asking them to compare and evaluate these, explicitly invites them to consider whether an association exists between the two marks. This approach is thus arguably leading, since without the invitation the question of whether the marks were produced by the same source might never have occurred to respondents.

300. As an alternative to these approaches, Simonson (1994) noted the growing use of experiments to test trademark confusion. Experimental approaches have several advantages as they enable researchers to develop scenarios that parallel actual buying situations while controlling the range of variables included. Perhaps more importantly, experimental studies provide greater levels of control and so enable researchers to isolate confusion from the allegedly deceptive source as opposed to confusion that may arise from other aspects of the product's appearance. For example, he suggested that if the brand name is in dispute, researchers need to be able to quantify the level of confusion that arose because respondents mistook one brand to be another. This confusion has to be distinguished from confusion that may have resulted from other attributes, such as similarities in the overall packaging design.

301. However, Simonson noted that experimental results may not be sufficient proof of confusion and he advocated investigating respondents' state of mind, so that the court has evidence of both behaviour and the beliefs upon which consumers acted. Furthermore, he suggested refining experimental approaches so that these contained a more detailed exploration of the attributes respondents associated with the brands involved in the dispute.

302. In particular, he proposed using a simulated choice approach, where respondents are presented with the junior and competing brands and asked to evaluate these on different dimensions. As part of this process, they would indicate their familiarity with the brands at issue, other products manufactured by the same company, advertising put out by the

brand, and so on. Depending on the associations respondents make, they are assessed as either confused or not confused. According to Simonson, two independent coders evaluate the responses, determine and initial classification, and resolve any differences that emerge through discussion to arrive at a consensus view.

303. Simonson argues that this approach simulates the choices respondents would make in real purchase situations. In particular, he suggested that its focus on the choices respondents make rather than on the specific trademarks makes it less susceptible to question wording or question order effects. As a result, he argued this approach could be used to test the extent to which respondents see the two marks as similar as well as any associations they make between the two marks and a single source.

304. However, while an experimental approach would enable researchers to address the issues of validity that affect other approaches Simonson has outlined, experimental studies are more complex and thus have not typically been adduced. Overall, while this approach offers some advantages, particularly, the greater level of control over the survey context, the lack of precedent increases the risk associated with this type of evidence.

305. Kapferer (1995) also examined the importance (and difficulty) of creating a valid survey context. He noted that judges have rejected survey evidence where they believed respondents had rationalised their answer in a way that would be unlikely to occur in a real purchase situation (see also Crespi, 1987). Kapferer suggested it is difficult for survey questions to capture the state of limited attention that applies to many purchase decisions and he argued that researchers should create this environment experimentally.

306. His proposed using a tachioscope to ensure a realistic context was created. This device projects an image of a brand or trademark (or other relevant stimulus) onto a screen for a very brief period time, after which respondents are asked to identify what they saw. According to Kapferer, this technique simulates the hurried and low-involved consumer scanning that typically occurs in fast-moving-consumer-goods (fmcg) purchases and offers insights into the likelihood of confusion that would occur in a typical retail context. He argued that as fmcg purchases are typically routine, consumers do not pay detailed attention to product labels, as surveys require them to do. Instead, they scan supermarket shelves and make rapid decisions based on recognition of visual stimuli such as brand names, pack colour, logos and other aspects of trade dress.

307. Kapferer reported an experiment involving 375 consumers who each saw four images exposed for one of six different time periods (ranging from $1/125^{th}$ of a second through to one second). Following each presentation, respondents were asked what they had seen,

what they thought they saw, and what they thought whatever it was they saw represented. All mentions of brand names were coded; the results indicated a wide range of different levels of confusion, and found that confusion levels differed considerably according to the time the brand image was exposed to respondents. Interestingly, the maximum levels of confusion were recorded at around 1/15th of a second exposure.

308. However, although this procedure offers an interesting advance on the traditional survey questions put to respondents, it does suffer from some limitations, which Kapferer outlined carefully. First, he noted that confusion may be exaggerated where the brand leader of a product category is one of the test brands. Respondents' tendency to identify the best known brand (i.e., the brand with the largest market penetration) is well known, and fits a general pattern of Double Jeopardy that applies to many different aspects of consumers' behaviour.

309. Kapferer also noted that respondents might use the name of a well-known brand in a generic sense, to refer to the product category and he outlined the need for additional research to examine how perceptual confusion (which occurs when one brand is mistaken for another) can be separated from generic confusion (where a well-known brand is used as a shortcut to describe a product category). He also found that the confusion levels documented relate to consumers' first encounter with a brand, and that subsequent encounters reduced confusion, especially if consumers were aware that two similar brands existed.

310. A further issue, which Kapferer did not discuss, is the length of exposure required in the tests. Although Kapferer tested several different exposure durations, it is clear that the level of confusion varied considerably and, after 1/25th of a second, the levels of confusion declined rapidly as the exposure time increased. Given that consumers pay varying levels of attention to a product category, any experiments using this methodology may require preliminary observational work to determine typical browsing behaviour, and the extent to which this varies by consumer type (regular purchaser within the product category, infrequent purchaser and new purchaser). It seems reasonable to assume that browsing time would be related to consumers' previous experience of the product category, thus analysing the results according to usage behaviour would be important to ensure any confusion was not simply the result of atypical exposure for a specific type of consumer. Senders and Green (1999) made a similar point when they noted that opportunities to deceive are created by limitations on human's information processing capacities.

Surveys of Confusion

311. Survey evidence has been adduced in several international cases involving allegations of misleading and deceptive behaviour. This section reviews some of the survey evidence from international cases that has received attention in the academic literature. Bottomley (2002) described a survey conducted in Hong Kong to test possible confusion between two crocodile emblems used on garments manufactured by Crocodile Garments Limited (CGL) and La Chemise Lacoste. His survey used a mall intercept technique to identify potential respondents in 17 locations throughout Hong Kong; respondents were shown emblems for five "control" brands and were also shown one of four test crocodile motifs, of which one was the Lacoste crocodile while the other three were images from the Crocodile Garment Company. Bottomley reported that 70% correctly identified the CGL logos as relating to CGL while 4% mistakenly thought they belonged to Lacoste. Of those respondents who viewed the Lacoste logo, 72% correctly identified the control brands while 22% identified the Lacoste image.

312. Respondents were also shown one of the CGL images next to the Lacoste image and were asked to identify each of these. Only a small proportion confused the CGL logo (62% correctly identified it while 7% said it was Lacoste); conversely, 26% correctly identified Lacoste while 47% said it was a CGL image. In addition, respondents were asked about the image they believed purchasers of each brand would have, and Bottomley also collected details of their own purchase behaviour. The findings revealed that respondents had quite different perceptions of the two brands and their own behaviour reflected these perceptions (the two stores had generally distinct consumer bases).

313. The survey was criticised largely for technical reasons: its use of a non-probability sample and the latitude interviewers had in selecting which individuals they would approach for an interview. In addition, the research was criticised because interviewers did not receive detailed written instructions and because verbatim comments made by respondents did not appear to have been clearly recorded.

314. However, counter-arguments that survey researchers were often constrained by time and budget requirements that limited their ability to follow pure statistical procedures were advanced as a justification for the use of non-probability sampling. Checks such as ensuring respondents' profile matched known population parameters were also documented and advanced as appropriate quality assurance procedures, and the survey was accepted.

315. As a result of this experience, Bottomley suggested a number of quality assurance procedures that could assist researchers to demonstrate the rigour of their survey estimates. Suggestions such as ensuring interviewers receive written training materials

and that all questionnaires are carefully signed and kept available so they can be provided in evidence if required, have been clearly established in New Zealand cases.

316. However, other suggestions contain new advice. For example, Bottomley suggested recording the sampling fraction interviewed during different day parts to ensure this could be used as a weighting variable if necessary, to address the fact that the profile of shoppers may vary by day part. He also recommended estimating the gender and age of every nth individual so that estimates for refusals can be compared against those of respondents in an attempt to measure potential non-response error.

317. In addition, he suggested ensuring that the sample comprises a very broad range of respondents. Thus, even if the researchers were primarily interested in one group of shoppers, Bottomley suggested that they consider drawing the sample from a wider population; this would enable comparisons to be made between those respondents who possessed detailed knowledge of the product category or brand with those who did not.

318. A second dispute that has been analysed in detail involved allegations that Kraft's claims about the calcium content of cheese slices were misleading and deceptive (*Kraft, Inc. v. FTC*, 970 F.2d 311 (7th Cir. 1992). Richards and Preston (1992) explained that the dispute arose because the FTC alleged Kraft's claims that the cheese slices they manufactured contained as much calcium as five ounces of milk were misleading. In actual fact, the FTC noted, Kraft's cheese slices contained only 70% of the calcium found in five ounces of milk because some of the calcium content was removed during the manufacturing process. In addition, the FTC challenged Kraft's claim that their cheese singles contained more calcium than imitation cheese products, a claim it argued was implied in Kraft's advertising.

319. To examine whether calcium content had a material effect on consumers' behaviour, Kraft commissioned a survey that examined two questions. First, whether consumers considered calcium to be an important factor in their decision to purchase Kraft Singles slices, and second, whether the difference between 70% and 100% calcium retention would affect their decision to purchase Kraft Singles slices or the way in which they used this product.

320. The survey used open-ended questions to examine why respondents bought cheese in general and Kraft Singles in particular. In response to the first question, just under 5% of respondents mentioned calcium and less than 2% cited calcium as a reason for purchasing Kraft Singles. Respondents were also provided with a list of seven ingredients and were asked to indicate whether Kraft Singles contained these ingredients (response options included yes, no and do not know). Eighty seven percent

indicated that Kraft did contain calcium and this group was then asked to indicate how important the calcium content was in their purchase decision. Over 70% of respondents indicated that calcium was very or extremely important to their purchase decision, and 95% indicated it was at least somewhat important. However, this question arguably favoured a finding of materiality, since three of the four response options offered suggested that calcium was important, at least to some degree.

321. Despite respondents attaching what appears to be a high level of importance to the calcium content of cheese, Jacoby and Syzbillo (1995) argued that this attribute was not material, since it was the second least important attribute of the eight they explored. However, Stewart (1995) argued that other attributes, such as "good tasting" and "having a real cheese flavour" were so self-evident that comparisons with them were meaningless. Thus instead of focussing on the relative level of importance, as Jacoby and Syzbillo did, Stewart noted that the key issue was whether the disputed attribute influenced consumers' behaviour.

322. Respondents to the Kraft survey were asked if they knew how much calcium was contained in a Kraft Singles slice; the vast majority were unable to answer this question. When advised that a Singles slice had 70% of the calcium contained in five ounces of milk and asked whether they would continue to purchase Kraft, 96% indicated that they would still purchase this brand, and, of this group, less than 2% indicated that they would use the product in a different way. Counsel for Kraft argued this result demonstrated that the calcium content of Kraft slices was not material.

323. However, the FTC interpreted consumers' lack of knowledge of recommended daily allowance of calcium as a sign that they would be more likely to rely on claims made in advertising. Given that earlier research by Kraft had established that the calcium RDA for girls aged 9 to 11 was 92mg, the FTC concluded that the difference in calcium content between a five ounce glass of milk and a Kraft cheese slice, which amounted to 60mg, was material. Thus, although Kraft argued these results provided strong evidence that the calcium content of the cheese slices had no material effect on consumers' purchase behaviour, the FTC rejected the survey findings.

324. The FTC advanced several reasons to support their decision. First, the FTC found that the range of options listed in the question examining how knowledge of the actual calcium content would affect consumers' behaviour was too limited. Instead of offering "yes", "no" or "don't know" options, the FTC argued that the survey ought to have included additional options that recognised respondents might reduce the rate at which they purchased the brand. Second, the FTC argued that although calcium was the second least important factor in consumers' purchase decisions, nearly three

quarters still described the calcium content as important or very important to them. The FTC held that even though other factors were more important to consumers, this did not mean that the calcium content was an unimportant factor in their purchase decisions. In addition, the FTC found that the survey did not expose respondents to the claims at issue in the case (since the interviewers were conducted by telephone, exposing respondents to the actual claims was difficult).

325. Stewart also criticised the use of a telephone methodology in the Kraft survey as: "*a poor vehicle for examining complex consumers decisions required by a test of materiality*" (p. 25). However, the FTC did not accept the criticism that respondents' inability to view the advertisements was a limitation even though this meant they were only aware of the specific disputed claim, and did not have this located in the context of the entire advertisement.

326. As well as considering the Kraft evidence, the FTC also submitted evidence of its own. Stewart (1995) outlined a copy test that he designed to examine the whether the advertisements in dispute communicated information about the calcium content of Kraft's singles and, if so, what information was communicated. The methodology involved exposing respondents to the advertisement claims and then examining their interpretation of these. Jacoby and Szybillo (1995) offer several criticisms of the FTC's evidence; these correspond to the sources of error discussed in Section 3.

327. First, the opposing researchers disputed the relevant universe from which the survey samples should have been drawn. Whereas Jacoby and Syzbillo (1995) argued that the relevant universe was people who had bought or who intended to buy individually wrapped cheese slices, Stewart argued that this definition was too narrow. He argued that the copy-testing Kraft conducted when developing and refining the advertisements in dispute used a much broader market, roughly corresponding to female homemakers aged 25 to 34. As a result, he claimed that the survey examining the potential for deception ought also to use a broader approach. Stewart suggested other milk users should also be included in the sample; for example, he identified people concerned about calcium deficiency, which could include pregnant or post-menopausal women. Furthermore, he suggested that any consumer who held concerns about the adequacy of their calcium intake, whether that consumer obtained calcium from milk or other sources, could also be considered a member of the relevant universe.

328. In the FTC's research, Stewart defined the relevant universe as women who were primarily responsible for their household's grocery shopping, who had children under the age of 18 living at home, and who had purchased cheese or cheese substitute products in the three months prior to the interview. He argued that this definition

recognised consumers who used the product in dispute, matched the definition Kraft themselves had of their target consumers, and ensured all respondents were exposed to advertising that contained either a disputed claim or another claim made in promotions for Kraft's Singles.

329. Stewart also defended the use of mall intercept surveys, although this sampling approach was not criticised by Jacoby and Syzbillo. However, like Bottomley, Stewart noted that although the courts have accepted mall intercept surveys, researchers should still check that samples obtained in this way are representative of the wider population.

330. Predictably, the main criticisms of the FTC survey involved issues of measurement error, specifically, whether the questions used were fair and appropriate measures of the legal question of interest. Jacoby and Syzbillo argued that the FTC surveys introduced a "yea-saying" or acquiescence bias, that the questions did not include explicit "don't know" options, and that the questionnaire lacked adequate control mechanisms.

331. Stewart noted that although questions can be designed to manipulate the type and range of responses provided by respondents, this is not the same as "yea-saying" bias, which occurs when respondents are more likely to agree with a statement because of the way it is presented or worded. While Stewart did not disagree that yea-saying bias may exist, he argued that its effects were likely to be small (since they are typically small in the surveys where this bias has been observed), and were therefore unlikely to detract from the estimates obtained or the conclusions based on these.

332. Stewart also questioned whether an explicit "don't know" option ought to be included and presented to respondents alongside substantive response options. He argued that a "no opinion" option can be presented in different ways ("don't know", "no opinion", "not sure"), and that these different forms are not necessarily synonymous. Furthermore, he argued that providing an explicit uncertainty response may encourage respondents to select this, even if they held a substantive view that they could articulate. Moreover, he suggested that, even where "don't know" or similar options have been provided, the overall response distribution did not vary significantly from that obtained when these options were not provided. In addition, Stewart noted that, even if the absence of a "don't know" option did affect the response distribution, this would also have affected the response distribution of the control sample. Since the main issue was whether the responses of the test and control groups differed, the over-riding criterion is that the two groups must have the same questions administered to them. That is, Stewart argued that the key issue was whether the questions themselves

were the same because researchers' interest was in the comparison of the test and control groups, not the absolute estimates obtained from each group.

333. Stewart also defended the questions used to explore respondents' reactions to the advertising they were shown. Respondents in the FTC survey were shown the Kraft advertisement together with a series of what Stewart described as clutter advertisements. They were then asked whether they remembered seeing an advertisement for cheese slices, what brand it was they saw advertised, and whether they recalled seeing an advertisement for Kraft Singles. Stewart argued that this question sequence is widely used in recall tests and he rejected arguments that his questionnaire should have included control questions that examined other milk attributes, such as vitamin content. However, he accepted that control questions exploring other variables could have strengthened the research design, although he argued that their absence did not invalidate the research.

334. In further rejecting Jacoby and Szybillo's arguments, Stewart claimed that the FTC research included several controls, since some respondent groups were not exposed to advertisements that contained calcium claims. Given that the research design aimed to compare responses from consumers who saw the disputed claim with those from respondents who had not been exposed to the claim, Stewart argued that an effective control had been built into the survey design.

335. However, Stewart noted the problem of controlling for pre-existing beliefs, since the Kraft advertisements had run for over two years before the FTC survey was conducted and had involved an advertising budget of over $15 million. This raised the question of whether, and how, researchers could separate out pre-existing beliefs about the brand from those that could be attributed to the advertising campaign. Respondents in the control sub-sample were exposed to Kraft advertisements that had run in the recent past, but that contained no statements about calcium content. Since respondents could also have seen advertising that included the calcium claim, it was possible they might volunteer that association following exposure to the control advertisement, even though the latter did not explicitly feature a calcium claim. As a result, Stewart argued that the survey estimates were likely to under-estimate the effect of the calcium claim on consumers' interpretation of the advertising claims.

336. The FTC survey contained both open-ended and closed questions; Jacoby and Szybillo were more critical of the closed questions. Stewart acknowledged the difficulty of designing closed questions that measure implied claims, since the questions need to be direct but non-leading, a difficult balance for researchers to attain. Stewart cited an earlier FTC judgment, where the judge stated that "*there is no way to test whether a*

consumer does or does not take a certain meaning from an ad other than putting that direct question to the consumer and asking the consumer to affirm or deny that the claim was made" (p. 22) (*Thomson Medical Co. v Federal Trade Commission 1984).*

337. As Stewart noted, there was no evidence that the use of an open-ended question in place of closed-questions would have resulted in a different response distribution. Furthermore, he noted that comparison of the test and control groups revealed marked differences in the response distributions and so reduced the likelihood that the question structure contributed to the pattern of responses obtained.

338. In defending the approach taken, Stewart noted the tradeoffs involved in developing experimental designs. Specifically, he explained the complexity of assessing materiality, given that a range of factors may influence consumers' behaviour. In Kraft, he suggested that testing whether the difference between the calcium content of 3.5 versus 5 ounces of milk is material may also be affected by knowledge of the level of calcium in Kraft and imitation slices, consumers' preferences for natural rather than imitation products, the price differential between natural and imitation products, and the assumptions relating to the cholesterol content of the two products.

339. Like Simonson (1994), Stewart suggested that some type of conjoint analysis may be the most appropriate method of ascertaining the relative importance of these factors. This approach uses a tightly defined experimental context and thus enables the role different attributes play to be estimated.

Confusion Threshold

340. Irrespective of the approach taken, the question of the extent of confusion must be assessed. Thus plaintiffs need to establish that a sufficient proportion of consumers would be confused about the source of the product at issue, or about the claim made in an advertisement.

341. Brient and Hebert (2000) suggested that typical trademark infringement cases had only to establish a 10% to 15% level of confusion between the trademark holder's brand and that of the junior user. However, they pointed out that many cases do not neatly follow this pattern and referred to *Hewlett Packard v Xerox* where Xerox Corp. began to make and sell recycled Hewlett Packard toner cartridges, used in computer printers. According to Brient and Hebert, Xerox's announcement of its intention resulted in an immediate response from Hewlett Packard, who claimed that Xerox was using its trademarks in a manner likely to create confusion. In particular, HP argued that consumers would be confused over whether HP sponsored or endorsed Xerox's

cartridges and it claimed infringement of specific trademarks that related to the recycled toner cartridges.

342. HP conducted a survey to submit in support of their request for an injunction to restrain Xerox's behaviour. While Brient and Hebert did not provide specific details of the survey design or implementation, they noted that of the 239 respondents surveyed, 24% identified HP as the cartridge manufacturer (as opposed to 66% who identified Xerox). In addition, 17% considered that HP had in some way endorsed Xerox to manufacture the cartridges.

343. However, although HP's survey appears to meet the 10% to 15% confusion threshold, the courts did not grant their request for an injunction. Rather than accept the estimates at face value, the court examined the reasons respondents provided for their answers. In some cases, respondents' assumption that HP had endorsed Xerox's manufacture of the cartridges arose because they noted the product was compatible with HP hardware. Ultimately, the court did not grant an injunction because it considered Xerox's right to provide information about its less expensive product, which benefited consumers, outweighed HP's rights as a trademark owner.

344. Harris (2002) suggested evidence that 15% of consumers are confused may be sufficient. However, he noted that as some respondents are routinely confused, estimates as low as 15% may not be sufficient to establish a likelihood of confusion, given that some confusion is likely to have arisen from sources beyond the product in dispute. The actual level of confusion plaintiffs are required to demonstrate varies from case to case as the level harm likely to result from confusion will also vary depending on the nature of the claims made and the product or service attributes.

345. Results of a telephone survey adduced in case T44/2002, an application by Amway New Zealand to register a device that included the words "PaintDirect", were not described in detail. However, Assistant Commissioner Hastie, described the survey as falling "*far short of what is required to ensure it is legally admissible*" and he took no note of it in reaching his decision. Nor did a follow-up survey conducted by email, which elicited responses from four of the companies that participated in the initial telephone survey, attract higher praise. The telephone survey was also criticised by an expert for the opponent (Wattyl New Zealand) who argued that PaintDirect could become uniquely identified with a single supplier in the same way as BankDIrect and PCDirect have become specifically associated with a specific source. In reaching his decision to permit registration, Assistant Commissioner Hastie also considered the sales and marketing activities undertaken by PaintDirect.

346. The question of what level of confusion needs to be established is also related to materiality, since evidence that a claim is leading consumers to make purchases they would not otherwise have made will demonstrate the material effect of the claim.

Assumption of Materiality

347. Interestingly, the FTC assumes that advertisers make claims because they wish to influence consumers' behaviour. This presumption of materiality essentially means that advertisers would have no logical basis for making claims that were unlikely to affect consumers' behaviour in some way. Richards and Preston (1992) argued that this presumption of materiality means advertisers seeking to defend themselves against FTC charges must be very careful to ensure their surveys contain a comprehensive list of response options that correspond to the full range of behaviours consumers may perform.

348. Richards and Preston also pointed out the difficulty of estimating materiality. Although the use of an importance scale in the Kraft survey represents one means of assessing both the absolute and relative importance of different product attributes, it is also open to criticism. As Richards and Preston noted: "*it requires no genius to expect consumers to rate virtually any attribute as important when asked*" (p. 52). As they go on to note, even asking the question implies the attributes investigated have some importance (else why would the researchers bother to explore consumers' responses to them). Furthermore, they suggested that consumers may be motivated to state that each attribute is important since failure to do so could result in the manufacturer removing that, and other unimportant attributes, which could reduce the overall quality of the product.

349. Richards and Preston also noted that courts' desire for survey evidence to be perfect, a task they describe as impossible given the inevitable trade offs researchers must make. Given the FTC's presumption of materiality, they argued that survey researchers face a difficult and demanding task, since survey evidence is bound to contain at least some debateable methodological decisions. In addition, they challenged the FTC's scepticism about advertising and the assumption that advertising claims would not be used unless they affected consumers' behaviour. As Richards and Preston noted: manufacturers rarely know "*why* an advertisement works; they only know (at best) *whether* it works" (p. 52).

350. Finally, Richards and Preston pointed out that although the FTC had identified flaws in the survey evidence provided by Kraft, it did not provide any guidance about the type of evidence it would find probative. As a result, Richards and Preston suggested that any evidence presented may be found to lack construct validity.

351. To address the absence of guiding principles, Richards and Preston presented a theoretical model that could be used to estimate materiality. This model examined the difference between a product's actual attributes and the beliefs held about these, and the relationship between these beliefs and purchase behaviour. Richards and Preston advised researchers to compare the effects of the deceptive claim with the effects of a truthful claim; if consumers' behaviour remained constant irrespective of the claim they were exposed to, the results would not support claims of materiality.

352. To implement their approach, they suggested researchers use a split-sample method where one group of respondents would view the allegedly deceptive claim while the second group would view the correct claim. Respondents would then be asked to rate the importance of the claim to their information search behaviour, purchase behaviour, and usage behaviours, using bipolar scales anchored by material at one end and immaterial at the other. Evidence in support of materiality would require that respondents' likelihood of behaviour was greater following exposure to the allegedly deceptive claim, and scores closer to the material than immaterial end of the scale (p. 54).

353. However, while this methodology provides guidance that is not outlined in FTC decisions, it is not immune from criticism. The methodology assumes respondents' stated intentions have a clear association with their actual behaviour, a claim that many consumer behaviour researchers have challenged. Furthermore, the bipolar scales proposed rely heavily on cognitive models of behaviour that researchers have argued may offer few insights into the routine behaviours that characterise most fast-moving-consumer-goods purchases (Ehrenberg, 1984). Thus, even a finding that respondents exposed to an allegedly deceptive claim rated an attribute as significantly more material than the group exposed to the true claim, does not mean that this difference will be evident in their purchase behaviour. It is well documented that what respondents say, and what they actually do, may differ greatly.

354. Although Richards and Preston's work represents a major advance in the development of a survey methodology that could aid researchers, it also has limitations that could reduce the value of evidence based on it. In particular, the lack of external validity makes the methodology vulnerable to criticism. Hoek and Gendall (2003) proposed an alternative methodology, which uses choice modelling to simulate a purchase situation and the alternatives consumers are likely to encounter. By pairing the allegedly deceptive attribute with other attributes, including the brand, it is possible to calculate the effect that attribute has on consumers' choice behaviour. While this measure is not

a measure of *actual* behaviour, it is nevertheless a closer approximation of this than cognitive outcome variables.

355. Examination of the *Lacoste* and *Kraft* cases highlights the need for researchers to control the sources of survey error outlined earlier. The remainder of this section considers additional issues that have arisen from these and other cases that researchers must also consider when designing surveys for use in intellectual property disputes.

Behavioural Correspondence

356. The validity of the survey context was highlighted in section 3 and has arguably become increasingly important as the range of disputes has increased. In particular, disputes over domain names and other properties that exist, or are typically used, in cyberspace has grown. In an early US case*, Simon Property Group LP v. mySimon*, Inc 104 F. Supp. 2d 1033 (S.D. Ind. 2000) reported by Harris (2002), the courts rejected a survey proposal because it would have failed to replicate consumers' thought processes as they encountered a disputed mark (p. 18). The proposed survey had planned to provide respondents with the homepage of two websites in sequence, but the courts held that search results would provide more information than the URL details, and that this would be material in determining which sites respondents decided to visit. In addition, the survey did not involve exposing respondents to search engine results, thus the court held that it could not provide an appropriate test of the likelihood of confusion between two domain names.

357. The complexity of consumers' purchase behaviours and the retail contexts that shape these has also pre-occupied researchers. US researchers have attempted to create realistic purchase situations by requesting respondents to buy particular brands during the course of their normal shopping expedition. However, they provided them with coupons featuring both the brands they had been asked to purchase and other brands, which may have primed respondents to select a particular brand, even if this is not what they were specifically asked to purchase. That is, respondents may have selected the brand they last saw or heard, which does not establish that they mistakenly selected that brand believing it to be another. Similar experiments have involved showing respondents a selection of brands, including one of the disputed brands, and asking them to purchase the brands they were shown during their shopping trip have also been criticised.

358. These approaches simulate shopping behaviour to some extent, although respondents' exposure to the brands they are asked to purchase would reduce the external validity of the experiment if it occurred immediately prior to purchase or involved higher levels of exposure than would typically be achieved through normal advertising media. In

addition, the instructions given to respondents would need to be very clear, so they understood they were to purchase the brands they were shown, and not simply a brand from the general product category. These experiments also require a control, since it is likely that at least some respondents would purchase brands in error, even where no confusion had been alleged. Overall, while these approaches introduce greater realism to respondents' task, the integrity of the experiments depends heavily on the level of exposure and respondents' understanding of what they have been asked to do.

A Joint Approach

359. Even where the survey has been competently designed and conducted, and there is clear evidence of the quality assurance procedures to attest to the processes followed, there is no guarantee that the survey findings will have a strong probative value. Of greater concern is the fact that there is no *a priori* manner of assessing whether the survey results will support the case advanced by those who commissioned and paid for the data collection. Because there are neither clear thresholds regarding the level of deception that needs to exist for a claim of misleading conduct to be upheld, nor a pre-defined level of association that needs to be attained before a particular attribute is considered distinctively associated with a given brand, lawyers and their research team inevitably take some risk when they commission and use survey evidence.

360. These factors complicate the use of surveys. However, contending with the fact that the survey will be subjected to detailed methodological scrutiny could be addressed if both parties conferred over the survey design, or if court approval of the design was sought prior to the data collection. There are several advantages to this approach. First, if the parties agree to the question wording, sampling procedures and data collection methods, the courts will have less need to be concerned with methodological details and will be able to focus attention on the survey findings and the extent to which these support the cases being advanced. As Sarel and Marmorstein (2002) pointed out, developing the perfect survey design may be an elusive goal. Instead, they recommended that researchers accept surveys inevitably suffer from some limitations; where these can be agreed upon, the cases may focus on more substantive issues.

361. A joint approach would also help address several of the criticisms that militate against the wider use of consumer survey evidence, namely that: it can be expensive, it is not always reliable, it may be open to multiple interpretations, and the results do not always support the side that commissioned the survey.

362. Surveys can certainly be expensive and forensic research will inevitably be more expensive than ordinary market research studies because of the higher standards required and the stricter quality assurance mechanisms that need to be put in place. It

will also be subject to a level of scrutiny most market research surveys never receive, and developing and implementing a design that will withstand this scrutiny is expensive. Agreement on a joint approach or design could help reduce the risk that the survey will be rejected on purely methodological grounds.

363. A jointly determined approach would also help ensure judges or hearing officers are presented with robust and accurate data that were collected by exemplary interviewers. However, even this does not guarantee that they will find the interpretation of the data plausible and compelling. Nevertheless, shifting the debate from one that focuses on methodological issues to one that examines how the data have been interpreted may increase the courts' willingness to accept survey evidence and the probative weight they attach to this. In turn, greater use of survey evidence will assist in the development of benchmarks against which future survey estimates can be assessed.

364. Sarel and Marmorstein's suggestion that the courts' emphasis has shifted to considering the weight that ought to be attached to survey evidence supports this reasoning. They argued that whereas the courts once dwelt on whether to accept survey evidence, evidence that the survey was competently conducted and reported now addresses the initial concerns that were raised. However, as Preston (1992) noted, the influence of survey evidence could be extended if the avoidable errors that plague surveys were eliminated. The following section examines Preston's criticisms in more detail; although his comments referred to US cases, the points he made apply equally well to issues that have attracted criticism in New Zealand judgments.

Flaws in Survey Evidence: Preston's Work

365. Although the criteria that "good" surveys should meet appear self-evident, implementing these is not always straightforward and researchers have faced many challenges in designing surveys that will withstand rigorous critical scrutiny. Indeed, even attempts to reduce or manage the errors outlined above have not prevented judges from rejecting consumer survey evidence or diminishing its probative weight. As a consequence, several researchers have noted that survey evidence is used to support rather than determine the outcome of a case (Jacoby and Szybillo, 1995), and have argued this may be as far as the influence of survey evidence can reasonably be expected to extend.

366. Preston (1992) had earlier anticipated this conclusion when he noted that the adversarial nature of legal systems means that even competently conducted surveys come under attack from counsel, whose role is to test the evidence adduced by their

opponents. Since surveys are never infallible, a varying level of challenge can usually be mounted to undermine a survey's authority. However, Preston's (1992) detailed overview of expert testimony and survey evidence research suggests that while competently conducted surveys can withstand minor criticisms, many surveys fail to influence the outcome of a case because of more fundamental (yet predictable) flaws.

367. He identified several flaws in evidence adduced in FTC and Lanham Act cases; although this evidence has been provided in a different jurisdiction, the fundamental problems raised are equally applicable to evidence presented in New Zealand hearings. Preston outlined three key flaws:

- A failure to ascertain how consumers interpreted the allegedly deceptive claim or practice. Consumers' interpretation of what a claim means tests whether a misleading and potentially deceptive claim has in fact been conveyed.
- Failure to ascertain materiality. That is, evidence has not tested the effect a claim has on consumers' actual or likely behaviour.
- Failure to establish the truth-status of the claim. A claim can only be misleading or deceptive if it makes a representation that cannot be factually supported by an objective examination of the product attributes.

368. Preston reported many instances of questions that have been rejected or deemed flawed. In some cases, open-ended questions have not been accepted because they were considered leading. In other cases, the questions were not sufficiently probed to elicit the full range of responses that could exist. Since the responses to open-ended questions must be classified and coded, the failure to supply a coding frame has also led to the rejection of some survey evidence. Even forced choice questions, which could overcome at least some of the difficulties associated with open-ended questions, have been rejected where their wording was found to be unbalanced, or where the range of response options provided was incomplete. Surveys have also been rejected where particular questions were withheld, or where the experimental design did not include proper control variables.

369. Interestingly, Preston's review suggests that the courts have been less concerned with issues of validity as they have rejected arguments that research conducted using natural conditions is more valid than research conducted in artificial conditions. Here, the courts have held that research must be designed to provide a legal test of conveyance or materiality, and such a test might be more appropriately administered in artificial conditions.

370. US court decisions have also emphasised the need for surveys to conduct fair tests, a point that has also emerged in a number of New Zealand judgments. Preston noted that surveys have been rejected where respondents were asked to use products in a way that was contrary to the product instructions, and where the attributes tested were not relevant to the issue at dispute, where the full range of competitors were not included in the experimental design, and where the stimulus material did not fairly represent the material consumers would see in the marketplace.

371. A common international problem is the failure to provide insights into the issue the judges must determine. Predictably, surveys not specifically designed to test the legal questions of interest have fared poorly in the courts. Preston reported that US courts have rejected image studies, brand penetration research and consumer product evaluations, on the grounds that these are not relevant to the dispute. These findings suggest that forensic surveys should be designed from first principles to ensure they are directly relevant to the dispute and specifically address the legal question of interest. As noted earlier, while other surveys can contribute important contextual information that provides those hearing the case with background details, these surveys were typically not designed to test the legal issue facing the courts.

372. Preston (1992) considered this latter issue was so self-evident that it was hard to understand why lawyers would use survey evidence when clear precedents existed to suggest that this evidence would not be accepted. He suggested that legal counsel would use the best available evidence, even if they knew this to be flawed, since it was not inevitable that those flaws would be identified and brought to the court's attention. When seeking to explain why experienced staff would behave in this way, Preston surmised that lawyers may feel under pressure to demonstrate to clients that they are progressing the case as staunchly as they can, or they may hope to introduce arguments not considered in earlier cases, that would lead the courts to accept evidence it had rejected on previous occasions.

373. He also noted the fact that survey researchers and lawyers know little of each other's craft: "*a researcher could make a wrong assumption about the law's requirements and hence recommend improper evidence, following which a lawyer, being ignorant of the research field and needing to rely on the researcher, accepts the recommendation*" (p. 64). Among other measures, Preston suggests that researchers should take more responsibility for their decisions, which he argues should not be made in order to advocate a particular position, but to test the strength and weaknesses of the competing perspectives. While this approach is important to ensure the validity of the research undertaken, it may also increase the risk of commissioning survey research, since lawyers

are instructed to represent their clients' positions, rather than determine the most elegant and realistic research design.

374. Preston recognised this conundrum and was sharply critical of the FTC for failing to provide advertisers and researchers with clear and straightforward guidance. He argued that the FTC's view of advertisers as adversaries prevented it from properly fulfilling its legal requirement to prevent deceptive behaviour. Clearer information about the evidence the FTC would find acceptable and probative would, Preston suggests, enable all parties to a dispute to use their resources more efficiently, and would also enable the FTA to allocate greater funding to preventative strategies.

375. The lack of defining principles has also arguably inhibited the development of forensic research in New Zealand since, without clear guidelines and criteria, the standards researchers are required to meet remain amorphous. We outline a series of criteria in Appendix 3; however, while these are gleaned from and based upon decisions issued in New Zealand, their use has yet to be tested by the courts. For this reason, the criteria are conservative and do not include the use of new methodologies, since there is no evidence of how the courts would respond to these.

6. Depth Interviews with Intellectual Property and Survey Research Experts

6.1 Methodology

376. To locate the review of recent cases and academic literature within a specifically New Zealand context, a series of depth interviews were conducted with senior law staff who had managed intellectual property disputes involving survey evidence, market research experts who had designed or conducted surveys used in intellectual property proceedings, and marketing academics, who had provided expert evidence in intellectual property disputes.

377. The interviews were conducted either in person or by telephone and ranged from 40 minutes to an hour and a half in length. A loosely structured interview protocol was used to guide the interview, although the nature of the discussion varied, depending on the role the respondent had played in IP disputes. A copy of the interview protocol is provided in Appendix 4.

378. The results are presented in three sections, since the issues raised by respondents differed. Section 6.4 provides an overview of the issues canvassed and highlights common concerns and points of difference among respondents.

6.2 Research Findings: Academic Expert Witnesses

6.2.1 Characteristics of Legal Surveys

379. Academics become involved when the lawyers dealing with a case believe that an expert perspective would add insights into the development or interpretation of survey research findings. Most report having an initial dialogue with the legal team to assess the role survey research findings could play, examine the availability of other evidence that might reduce the need for primary consumer research, and discuss the risk that the research results might be unhelpful (and so outweigh any benefits that could be gained from undertaking a survey). Experts found they could contribute a great deal more to a case when they were involved from the outset, and contributed to the design of a survey. Where they were asked to comment on survey results without having had the opportunity to influence the design of the survey, they found their role was more complicated, particularly if the survey contained avoidable flaws.

380. Although experienced researchers, academics noted the importance of clearly establishing the legal question of interest and ensuring that any research undertaken addressed this directly. However, while they knew the research objectives must provide evidence a judge or hearing officer would find helpful, they were aware their background did not always enable them to understand the legal nuances of a case. Some also commented that, while lawyers were clearly experts in their field, they did not always understand the limitations of survey evidence, what questions surveys could actually address, or the trade-offs necessary in survey design. These comments foreshadowed a theme that emerged several times during the depth interviews: the tightly defined knowledge held by different members of the team and the importance of ensuring that knowledge (and the limitations of knowledge) was shared among the team members.

381. Although the academics with experience in this field were familiar with sophisticated research methodologies, their general rule of thumb was to keep the research design and data analyses as simple as possible. There was a strong awareness that, although the research needed to be very robust, it had to be designed to persuade an intelligent lay audience. Thus while the research would be scrutinised by expert peers (acting for opposing counsel), the findings must be straightforward and compelling, so they could be easily understood by those without formal training in research methods. This guiding principle emerged in several comments made by respondents who had provided expert evidence in intellectual property cases. In practical terms, academics attempted to use the simplest analyses to illustrate their arguments. For example, they may comment on practical rather than statistical significance.

382. As well as recognising the need for the survey results to be accessible to a lay audience, academic experts also noted the importance of ensuring the survey met higher standards in both its design and implementation. Standards deemed suitable for commercial surveys were considered too low for material to be presented in court. Some academics commented that not all research companies recognised that surveys adduced in court would be subject to more detailed and critical scrutiny than standard market research reports. However, they were very aware that failure to recognise the higher standards required could result in evidence that was easily attacked and undermined, and that could ultimately have little influence on the outcome of the case.

383. Academic researchers also noted that surveys must have strong external validity and that the tasks respondents were asked to perform should replicate (or at least simulate) the purchase situations they may encounter. However, they also noted the difficulty of obtaining high external validity, particularly given the nature of some purchase situations, which could require respondents to be intercepted in store or near point-of-purchase displays. While desirable to conduct interviews in as natural a setting as possible, academic researchers also pointed to the multiple distractions present in most retail environments and the difficulty of controlling confounding variables. Given these problems, they noted the importance of pragmatic compromises that recognised the ideal circumstances in which the research would take place, and those researchers were able to achieve within time and cist constraints.

384. More specifically, respondents were asked to comment on sources of survey error, the importance they attached to these and the mechanisms they used to ameliorate the effects of different types of error. The sections below correspond to the sections in part 4 of this report.

6.2.2 Coverage Error

385. All respondents noted that the relevant population was defined by the issue that was to be determined. Thus, if the dispute involved a particular product category, then consumers who purchase from that category were typically the population of interest. However, respondents also noted the importance of ensuring that the population extended to both current and potential users, to ensure that views from people at actual and potential risk of deception were canvassed. The relevant population was defined in consultation with the legal team heading the case, and with the market researchers who would be responsible for the data collection. This collective approach was seen as critical to ensuring the population could be adequately sampled (i.e., was practical and accessible), and that it would meet the legal needs of the case.

386. Academics also spoke of the need to consider multiple populations in some cases. While consumers were often considered initially, some cases also required an assessment of confusion that could occur among trade members, and separate and quite distinct surveys were required to assess this question. For example, cases that involved the use of point of sale material, or decisions where consumers normally sought advice from retail staff before making their purchase decision, may require an exploration of retail staff members' understanding of the disputed claims or attributes.

387. Academics also noted the importance of ensuring the sample matched the population of interest. However, while samples and populations can be matched on known characteristics, such as age, gender and occasionally other demographic traits, it is often not possible to match them on other attributes, which may be strongly related to the variables of interest.

388. Academic experts reported using screening questions to assist them identify relevant respondents from a wider cross-section of the population. In general, they felt that well-designed screening questions that did not alert respondents to the legal issue were valuable and could prevent considerable wastage by ensuring respondents with no experience in the product category were not interviewed.

389. Where mall intercept surveys are undertaken, respondents noted that the geo-demographic profile of mall shoppers could be obtained from the Department of Statistics and that this could provide a useful benchmark against which characteristics from the sample could be assessed. However, even this profile can only ensure that shoppers match the wider population on demographic attributes, and these may be only weakly correlated with the variables of interest.

390. Because mall intercept surveys presented difficulties in matching the sample against a wider population, some academic researchers indicated that they would be reluctant to use a mall intercept sample. However, those with this view suggested that if the interviews were conducted across several cities, the problem of matching the achieved sample with a wider population could be mitigated. Overall, most respondents indicated that they thought mall intercept samples would be acceptable, although they noted the need to use multiple sites so that that the resulting estimates could be extrapolated to the wider population.

6.2.3 Sampling Error

391. Academic respondents were very conscious of the cost of research fieldwork and the effect on cost that even quite small increases in the sample size would have. Their comments revealed a tension between ensuring the sample would provide adequate cell sizes for more detailed analyses, on the one hand, and remaining within a budget that clients had set, on the other.

392. Because the sample size is dependent on the types of analyses required and the extent to which the population comprises sub-groups whose responses are also of interest, academic respondents were reluctant to suggest an "optimal" sample size. However, some reported surveys based on samples of 300 respondents and, despite the fact that some IPONZ decisions had made adverse comment about samples this size, respondents did not recall any criticisms having been levelled at samples of 300. Where possible, most opted for larger sample sizes, typically at least 500 respondents, which they felt provided precise estimates without dramatically increasing the overall cost of the fieldwork.

393. Several respondents also commented on the need to ensure that the overall sample was large enough to support any sub-sample analyses that might need to be undertaken. In some cases, this may require a larger overall sample size, depending on the size of the groups that would be used in subsequent analyses. Overall, respondents recommended that the sub-samples should contain at least 100 respondents, and noted that the maximum error margins associated with analyses based on a group this size would be nearly 10%.

394. Overall, respondents were less concerned about the sample size, possibly because where differences in the estimates obtained were large, the precision of the estimates (the size of the confidence intervals) was less important. Clearly, where respondents' interpretations were more evenly matched, more attention may be paid to the size of the sample and the extent to which observed differences were "real".

395. The lack of clear guidelines about benchmark levels of confusion or distinctiveness caused some concern to academic experts. They noted that more information about how judges interpreted and assessed survey estimates could assist them to determine appropriate sample sizes. In particular, this knowledge could assist them to interpret results from pilot studies and assess the viability of extending these to full-scale surveys.

396. For researchers, establishing whether the survey is likely to yield large differences may be very difficult to predict, particularly where there is little market information available. In these cases, pilot surveys may be useful as these can establish the level of

association or confusion among the population of interest and may inform the decision of whether to proceed with a larger scale project. However, questions remain about the status of pilot surveys and the extent to which these may need to be disclosed to opposing counsel. Clearly, pilot surveys will do little to reduce the risk associated with undertaking a larger scale project if the findings from preliminary work are unhelpful and non-confidential.

6.2.4 Non-Response Error

397. Academics acknowledge that response rates had fallen and that it could be problematic achieving response rates of over 50% (the minimum recommended by the United States Judicial Manual). While samples could be drawn until particular demographic quotas were filled, this did not disguise the problem that the overall response rate could be very low. Although some respondents felt that the response rate needed to be considered as part of an overall assessment of a survey's robustness, there was a general feeling that surveys where the response rate fell below 35% to 40% were problematic and difficult to deal with, and some considered 35% was the minimum acceptable response rate.

398. Respondents noted that commercial telephone surveys sometimes had response rates as low as 10%, a figure that leaves a large potential for non-response error. They also acknowledged that, while the achieved sample could be checked against population demographic traits, these may not be correlated with the variables of interest. As a result, non-response error remained very difficult to quantify and address.

399. One proposed method of addressing non-response would involve seeking follow up interviews with non-respondents. However, while the expert respondents noted that this would be ideal, follow-up surveys could be difficult to achieve in practice. In particular, pursuing respondents who had already indicated their unwillingness to participate in a survey could raise ethical issues that would need to be balanced against the desirability of ensuring the sample was representative of a wider population.

400. Some respondents suggested that they would be willing to accept a trade-off between the response rate and quality of interviewing, and suggested that if they could guarantee the interviewers had performed well, this could mitigate a lower response rate. However, high performing interviewers tend to achieve better response rates, thus the proposed trade-off does not appear to reflect interviewers' experiences.

401. Respondents also noted that, because research companies were charged with conducting survey fieldwork, staff at these companies were also responsible for calculating the survey response rate. Experts noted that they did not always have access

to the particular formula used. Given the wide range of response rate formulae that may be used, these comments imply that different companies may use different formulae; if this is the case, comparisons between response rates may be flawed. The American Association of Public Opinion Research has developed several formulae that may be applied to different survey modes and that specify the assumptions made about different categories of respondent. Adoption of standard procedures from established guidelines such as these could help address differences in research companies' practices and ensure that consistent formulae were applied by all parties. While some respondents supported using more liberal formulae (for example, assuming a proportion of ineligible respondents among those who could not be contacted) others preferred a very rigorous and more conservative formula (for example, assuming all uncontacted respondents would have been eligible to have participated in the survey).

402. Irrespective of which approach would be used, the adoption of a consistent approach would enhance the comparability of response rate calculations. However, experts suggested that judges and hearing officers typically paid little attention to response rates and rarely commented on whether non-response error undermined the validity of the results obtained. As a result, although experts recognised the value of employing uniform formulae, they questioned whether standardising the reporting of response rates reported would have a material effect on how judges or hearing officers viewed the survey evidence.

6.2.5 Measurement Error –Questionnaire Design

403. Academic respondents recognised the tension between open-ended questions, which are non-leading in their design, but that can be difficult to code, and closed questions, which are much easier to code, but that may be challenged for being leading or failing to provide a full range of responses. All respondents commented on the danger of using leading questions, but suggested that while some leading questions were clearly identifiable, the concept of being leading was a continuum rather than a dichotomy. Thus, while obviously leading questions were straightforward to recognise, there was also a grey area where elements of a question or a question sequence could be seen as leading.

404. In particular, academic experts commented on the need for a sequence of questions that eventually directed survey respondents to the topic of interest. Here, they considered that the design of this sequence was a "craft" rather than a science, and the sequence depended very much on the legal issue being investigated and was probably not amenable to a "rule-based" approach.

405. Nevertheless, respondents agreed that the sequence should follow a movement from general to specific questions and that this "funnelling" approach should eventually lead the survey participants to the issues at dispute. They also recognised that the initial general questions used to introduce survey participants to a topic often lacked context and so elicited a wide range of responses, many of which were unhelpful because they were unfocussed. Yet while they agreed that diverse responses could prove damaging, since they could suggest a lack of relationship between the attribute examined and a particular product category or brand, they were unable to suggest an alternative approach that would not be at risk of leading respondents. In general, academic experts agreed that some direction could be provided to survey participants, but that direct introduction of the brands in question, or even the sub-category of interest, would run the risk of being viewed as leading.

406. Some respondents had used closed questions, where survey participants are presented with particular response options (and usually a "Don't know" and "Other response" category). They reported that these had also worked well, but found that they were most useful when employed to examine survey participants' behaviour or knowledge, and where the response options could be easily anticipated.

407. Where closed questions were used, academic experts noted the importance of demonstrating that rigorous pre-testing had been undertaken to ensure the completeness and appropriateness of the categories used in the question. They felt that, where sound pre-testing had been undertaken, closed questions were more rigorous and, in their opinion, were preferable to open questions, which inevitably involved the subjective judgment of researchers.

408. Although the academic experts interviewed did not rely on a structured process for developing questions or ordering the sequence in which these were presented to survey respondents, all noted that questionnaire development was a critical phase of the research project. They reported going through several iterations of the questionnaire with the research company, and seeking peer review as a quality assurance mechanism. Some academic respondents who had experience as practising market researchers noted the importance of "group mind" reviews where they drew on the expertise of colleagues.

409. Experts also recognised the importance of pre-testing the questionnaire to ensure that serious issues in respondent understanding and question wording were identified and corrected. They noted that pre-testing typically used a convenience sample and did not employ cognitive evaluations of the questionnaire (such as "double-back" tests where respondents explain the question they thought they were answering). As a result, while this type of pre-testing may identify errors in the question sequence, problems in the

instructions given to interviewers, and lack of understanding of particular words, it is not usually designed to test whether respondents' overall understanding of the questions parallels researchers' intended meaning.

410. The actual responsibility for designing and reviewing questions appeared to rest with the research company, rather than with academic experts, who typically oversaw the research process and provided feedback, rather than directing day-to-day operational decisions. Although the experts interviewed had good relationships with the research companies they had worked with, the specific role they played in each project appeared rather ad hoc and depended to some extent on the expertise and experience of the project director in the research company.

411. Although experts had opinions about the relevant benchmark levels of confusion or distinctiveness that should be demonstrated, they felt there was little clear guidance from the courts about what was considered misleading and deceptive. While they appreciated that this definition would depend on the nature of the product and consumers' own knowledge, they nevertheless felt some statements about expected levels of confusion or distinctiveness could assist survey interpretation.

6.2.6 Measurement Error – Interviewer Quality Assurance

412. Practices with interviewers tended to vary. Where possible, academic experts would attend interviewer briefing sessions, although they did not typically play an active role in these. Views on the desirability of providing video evidence of training sessions differed; while some thought this evidence could attest to the quality procedures undertaken prior to interviewing, others thought that the expense of videoing each training session could outweigh any benefits that might be expected to accrue. Others also noted that while videos of perfect interviews in training would provide some evidence of the overall quality of the data, the interviewers' behaviour might differ in the field, making random spot checks of the interviewing important. However, all agreed that where face-to-face interviews were being conducted in several centres, evidence of the training sessions held in each centre was useful to demonstrate that comparable standards of interviewing had been maintained throughout the data collection.

413. Academic experts relied on the research company to recruit and train interviewers and noted that some research companies had sought and obtained international accreditation that demonstrated they had met particular standards for interviewing quality. The interviewers' credentials were considered very important, and academic researchers expected that market research companies would use only their most experienced and able interviewers to collect data that may be adduced in court proceedings. However, they also noted the difficulty of assessing the quality of

interviewers' work and commented that errors in this were often very difficult to detect, particularly once the data had been entered into a numeric format.

414. Some respondents reported having compared tape-recorded interviews with written interview transcripts and agreed that there tended to be discrepancies between the two versions of the survey. In particular, interviewers tended to summarise respondents' answers in order to be able to cope with the speed at which answers were provided. However, academic experts recognised that interviewers needed to adopt a "pragmatic" approach to transcribing answers and did not necessarily consider that discrepancies between oral and written versions of an interview were problematic.

415. Yet while academic experts recognised that tape-recorders could simplify interviewers' task, they also noted that any discrepancies in interviewers' conduct would be available to opposing counsel. In particular, they noted the need for interviewers to remain completely neutral throughout the survey and stressed the importance of ensuring interviewers were not advised how the results would be used or who the survey client was. Where interviewers were aware of the survey topic, there had been instances of over-zealous probing to obtain particular responses and prompting where this was neither required nor allowed by the survey instructions.

416. Those experts who had discreetly observed interviewers in the field were generally satisfied with the quality of work they viewed. However, they noted that interviewers could occasionally struggle to fill particular quotas (where the area in which they were operating had a specific demographic profile) and suggested that some discretion was appropriate to enable "pragmatic" judgments to be made. This comment highlights the tension between urgent practical issues that might arise in the field (despite pre-testing) and the need to ensure survey instructions are followed to the letter to avoid any challenges that might subsequently be made to the quality of the data.

417. Academic experts' views on this topic suggest that they trusted interviewers to make sound judgments about how best to convey the main points a respondent was communicating to them, or to deal with sampling issues, should these arise. However, where interviewers interpret rather than record respondents' comments, they risk reporting inaccurate survey responses and, if cross-examined on this issue, they could call into question the validity of the answers recorded.

418. The nature of the questions used in surveys and the level of autonomy conferred on interviewers have a critical influence on the quality and integrity of the data collected. At present, there are several tensions that affect the design and implementation of surveys. While open questions are more likely to satisfy concerns judges and hearing

officers have over the neutrality of the approach taken, this style of question may require interviewers to exercise more judgment over the way in which they interpret and record survey participants' responses. Similarly, where sampling or question wording issues arise during the interviewing, allowing interviewers to respond to these without reference to research staff may create opportunities for the integrity of the data to be challenged.

419. Given the concerns that could be raised about the interviewers' potential influence on the responses recorded, more detailed quality assurance, including field observations, matching taped and written transcripts, and detailed training would seem prudent. However, given some academic experts' experiences with taped interviews, routine use of this procedure as a quality assurance mechanism implies that interviewers' behaviour must reach much higher standards if it is to avoid becoming an Achilles' heel that undermines the credibility of the evidence.

420. Additional measures could include on-going visits from the project director; this would ensure responsibility for any changes to the survey wording or sampling were made by those with expertise in question wording and sampling. Clarifying responsibility for decisions such as these could reduce the survey's vulnerability to allegations that changes were made expediently and would ensure that interviewers were not cross-examined on matters in which they were less likely to have expertise. However, while these measures would increase the integrity of the data, they would also increase the overall costs of the survey, a point that may not be acceptable to lawyers or their clients.

6.2.7 Role of Experts

421. Experts were also asked to consider how their role could develop in the future. At present, both parties to a dispute often engage an expert to assist with the design, interpretation and evaluation of survey evidence. Those involved in the design support and defend the approach taken while those charged with evaluating this approach identify the assumptions made and the logic of these, and highlight any errors made in designing, collecting, analysing or reporting the survey findings. While this production and testing of evidence ensures those charged with determining a case can make a robust assessment of both the merits and limitations of survey evidence, where experts hold distinctly different views, it can be difficult to choose between these. Faced with this dilemma, judges and hearing officers may instead elect to rely more heavily on their own experience as consumers, or on other, less disputed, evidence that is available to them.

422. To address an impasse where expert evidence presents opposing arguments, academic experts were asked to comment on a proposal that would see a court appointed expert assist the judge to evaluate other experts' submissions and reach a view about these. There was strong support for this suggestion, although respondents also noted that the

role of a court-appointed expert should be to resolve technical issues, and not to provide commentary on the quality of the individual briefs that had been submitted. The latter responsibility would remain with the judge or hearing officer, although it would seem likely to be influenced by the quality of the technical detail these briefs contained. Nevertheless, respondents considered that a court-appointed expert could enable a more considered approach to be taken to survey evidence and would allow an expert party seen as completely disinterested to provide guidance to those charged with determining the case.

423. Respondents were also asked to comment on the proposal that jointly commissioned surveys be undertaken to reduce the likelihood that both parties to a dispute would design and commission survey evidence and adduce findings from different questions that, at least in some circumstances, could support quite different conclusions. Again, there was strong support for a proposal that could reduce disputes over methodological issues and move discussion towards debate over the interpretation of the research findings.

424. Some respondents commented on the merits of an inquisitorial rather than adversarial approach and suggested that the former could be more likely to test the quality of the evidence and facilitate a detailed discussion of the actual survey results. Experts felt that this initiative would be likely to improve, rather than detract from, the overall evaluation of survey evidence.

425. However, others noted that, although they saw advantages from their own perspective, they felt lawyers may feel less inclined to recognise these. In particular, they suggested that lawyers who felt they had retained the most competent research company and the most persuasive experts could feel reluctant to give up these advantages.

426. Experts were also asked about the possible use of more sophisticated methodologies, particularly those that could simulate behaviour and thus provide insights into what respondents may do in particular scenarios. At present, research methodologies focus almost exclusively on respondents' opinions, which are more open to interpretation and thus dispute. However, while they agreed that more sophisticated methodologies could provide better and more detailed insights into cases, they were concerned that the technical nature of the analysis undertaken could prove difficult to interpret and so may reduce the importance attached to the evidence.

427. Several spoke strongly against the use of analyses that were more sophisticated than bi-variate comparisons, and suggested that, if clear patterns were not evidence in simple

analyses, more sophisticated procedures would be unlikely to uncover these. For this reason, these respondents questioned the use of modelling and suggested that simple procedures and analyses would be more accessible and, as a result, more convincing.

428. While the appointment of a court expert could ensure more technical evidence was made accessible to judges and hearing officers, respondents felt the courts would continue to prefer more straightforward evidence. As a result, they believed that while modelling and other approaches could provide insightful results that directly informed the legal question of interest, they were concerned that the complexity of this evidence made it riskier and reduced its potential probative value.

6.3 Research Findings: Market Researchers

429. Market researchers and experts often work closely on the design of a survey. However, market research companies are responsible for collecting the data and typically manage this aspect of the research process. Because market researchers and academics have similar knowledge, the same general interview protocol was used with both groups.

6.3.1 Coverage Error

430. Defining the survey population was seen as critical by market researchers and their comments were very similar to those made by academic respondents. Both groups relied on the legal team to define the scope of the legal question and their interpretation of the relevant population depended heavily on the response to this issue. Because they were responsible for collecting the data, market researchers tended to be more concerned with the availability of relevant sampling frames and the practical questions of accessing a specific population without incurring wastage.

431. The main way they ensured they could reach particular populations was to conduct the survey in an area where they could access a wide population cross-section, and then to employ screening questions that enabled them to select individuals whose characteristics meant they belonged to the population of interest. However, while screening questions reduced wastage, researchers noted that it was important that the screening questions did not alert respondents to the survey topic, and thus sensitise them to subsequent questions.

432. Most market researchers reported using age-gender quotas, particularly where the population was very general. However, some noted the use of age-gender quotas even when the population of interest might be expected to have a particular demographic skew. Where demographic traits are not clearly related to the variable of interest, the

continued use of age-gender quotas may provide opportunities for challenges to be levelled at the match between the achieved sample and population.

6.3.2 Sampling Error

433. Market research experts consider that judges and hearing officers place a strong emphasis on sampling error, thus they tend to view samples containing fewer than 300 respondents as open to challenge (the maximum error margin for a sample of 300 is 5.7%).

434. The sampling procedures employed tend to depend on the survey mode. Where visual stimuli are not required, most surveys are conducted by telephone, where random digit dialling ensures that the sample is a simple random sample. However, where showcards or other visual stimuli are required, surveys are conducted in shopping malls or in respondents' homes.

435. Face to face surveys that take place in respondents' homes usually involve clustered samples, where a series of households are contacted around a randomly selected starting point (this reduces the travel time and cost of the survey). However, clustered samples are less efficient than simple random samples and the design effect is greater than one. As a result, the sample estimates are thus less precise and the error margins increase (the actual size of the increase depends on the size of the clusters). Although clustered samples are not simple random samples, few respondents commented on using the design factor to estimate the increase in error margins.

436. As academic experts noted, if the proportion of people who hold a mistaken belief, accept an incorrect attribution, or make a particular association, is large, the size of the error margins becomes less important, and this may explain why market research experts do not integrate the design effect into their error calculations. However, in cases where the estimates suggest some equivocation among the relevant public, incorrectly calculated or large error margins may reduce the confidence placed in the survey findings. Ensuring that the response rate calculation corresponds correctly to the sampling procedure employed will reduce opportunities for technical criticisms to be levelled at survey evidence.

437. Mall intercept samples have increased in popularity as they represent considerable cost savings over at-home face-to-face interviews. Technically, they are not equivalent to simple random samples although they have been viewed as such in international judgments. As academic experts noted, knowledge of mall traffic profiles could be compared to specific population characteristics, thus enabling mall intercept samples to be checked against known parameters.

438. However, the rigour of mall intercept samples could be improved. Some market researchers noted that they did not require the use of randomising procedures to identify eligible respondents. Without selection procedures in place, interviewers may approach people who they think look helpful and more likely to participate, and avoid those who look intimidating or unappealing. These often unconscious selection biases can produce a seriously biased sample.

439. Yet while market researchers did not always require interviewers to use standard selection procedures, they were very aware of the need to balance the sample across different day parts and days of the week. Their comments reflected a strong awareness that the sample should be fairly chosen from the wider population of mall shoppers and the use of full day and week interviewing would reduce challenges about whether the achieved sample reflected the overall range of people passing through a given shopping centre.

440. Although researchers recognise the desirability of interviewing over several days and day parts, they also noted that practical considerations, particularly the urgency with which results were required, sometimes meant compromises were necessary. As a result, a number of surveys are conducted over malls' busiest days, in the hope that shoppers passing through on these days approximated the wider group that visited that mall.

441. While researchers accepted that these compromises could skew the sample, they typically replied on post-hoc weighting to reduce any imbalances detected. As noted earlier, while weighting can correct samples to ensure they match a population on known characteristics, those traits may not be related to the variable being estimated. If this is the case, both the logic and likely benefit of weighting are unclear, and the decision to weight may be challenged by experts assisting the opposing party.

442. Some survey research experts suggested that web-based surveys could become an important survey medium for forensic research as they enabled a wide array of visuals to be used in addition to a high level of simulation. However, until the sampling frames for internet surveys become more widely available and robust, this medium will be open to challenge on methodological grounds.

6.3.3 Non-Response Error

443. Market research experts were acutely aware that survey response rates had fallen and that this trend created difficulties when surveys were adduced as evidence. Although response rates for face-to-face surveys remained around 50% (some reported face-to-face

response rates of 60%), those for telephone surveys were typically only around 20%, and sometimes lower.

444. Market researchers considered that the minimum acceptable response rate was 50%, although they felt that the minimum depended to some extent on the sample and the ease with which respondents could be recruited. However, where response rates dropped below 40%, market research experts suggested some form of non-response follow-up should be undertaken, although they conceded that time and cost constraints meant this rarely occurred. Furthermore, although differences between respondents and non-respondents could be recorded for some variables, such as age and gender, these observations may not provide any insights into the behaviours of interest. Thus, while some follow-up of non-respondents was desirable, market researchers found it difficult to suggest how this might occur.

445. In addition, market research experts noted that, while low response rates increased the potential for non-response error, a low response rate did not mean this was inevitable, although this argument became increasingly difficult to sustain as the response rate decreased. Overall, they agreed that the higher the response rate the better, although they felt that low response rates had not seriously compromised surveys in which they had been involved.

6.3.4 Measurement Error –Questionnaire Design

446. Market researchers also recognised the desirability of using open-ended questions, although they noted that the development of coding frames to classify responses to these could be open to challenge. Some had developed classified open-ended questions only to find that opposing counsel commissioned another researcher to prepare an alternative coding frame.

447. Most devolved the task of preparing coding frames to staff who were members of the project team, although few reported undertaking any verification process, such as employing two coders and providing evidence of the agreement between them. Given that academic research involving qualitative evidence, such as responses to open-ended questions, is almost invariably subject to inter-coder reliability tests, it is interesting to note that this practice does not appear to have extended to commercial market research.

448. Since competing frameworks have been used to classify the same open-ended data, coder validity checks could strengthen the conclusions based on open-ended data. If a case relies heavily on findings from qualitative surveys, coding checks would seem even more important, as challenges to the coding frame developed could potentially undermine the role any survey evidence could play. Furthermore, verification of the

coding frame could reduce challenges that opposing counsel may level at the classifications developed and the conclusions based on these.

6.3.5 Measurement Error – Interviewer Quality Assurance

449. Not all market researchers involved in forensic research managed a field force and at least some contracted this task to other market research companies. When selecting these, they considered that accreditation provided evidence that the interviewers were well-trained and their standard of work could be relied upon.

450. Where field work was sub-contracted, market researchers did not always brief the interviewers, although they may have assisted with the preparation of notes used in the interviewing. The level of scrutiny interviewers came under also varied – while some market researchers undertook site visits during the interviewing, others did not observe the interviewers at work and left the responsibility for this to the field force supervisor.

451. Market researchers noted that it was routine for at least 10% of interviewers' work to be audited, although some were unclear about whether the data quality was audited or whether the supervisors simply verified that an interview had taken place. Given the scrutiny applied to survey questionnaires, it would seem prudent to ensure that the verification process was clearly documented and that all members of the research team were aware of what the audit reviewed.

452. Some market researchers had been involved in surveys where interviewers had tape recorded the interviews they conducted. This was not an initiative they had suggested, but a response to a request from the legal team responsible for commissioning the survey. While market researchers recognised that tape transcripts that clearly matched the written questionnaires completed by interviewers would enhance the credibility of the data, they also noted that any errors the interviewers made were immediately obvious and that tape-recordings could prove unhelpful.

453. Even where they had achieved high standards of quality control and had controlled the effects the errors outlined above could have on the survey estimates, market researchers were not sure what benchmarks had been established in deception or distinctiveness cases. Knowledge of typical standards would be very helpful, they suggested, in determining whether results from pilot surveys supported the development of a full-scale survey.

454. However, while some felt that the levels of association or confusion that were required were often not clear, they also recognised the difficulty of establishing clear benchmarks, noting that these depended on the product category involved, the likely risk

consumers faced, and the consequences, should deception occur. Yet although relevant standards might vary from one case to another, more detailed communication between the legal team and the research group could help clarify this question for specific cases.

6.4 Research Findings: Lawyers

6.4.1 Role of Survey Research

455. While both academics and market researchers were enthusiastic about the potential role survey evidence could play in intellectual property disputes, lawyers were more cautious and placed greater emphasis on the risks associated with survey evidence, and the cost of commissioning a survey. In particular, they believed that rigorous surveys were difficult to design and that other forms of evidence may not carry such a high risk. For this reason, several noted that the use of surveys had not increased over recent years, and commented that they did not expect to see this trend change in the near future.

456. Others felt surveys were riskier forms of evidence because this type of evidence was not always well-understood by judges and hearing officers. They questioned the level of knowledge those hearing cases had of survey evidence and some suggested that, because they perceived a "lack of sophistication", survey evidence carried greater risks than other types of evidence. These respondents suggested that uncertainty over how survey evidence would be received and interpreted, together with the costs of commissioning surveys, had reduced lawyers' willingness to rely on this type of evidence.

457. In addition, some commented that where the issue in a dispute revolved around public perception, surveys may not provide information that the courts found useful. Respondents holding this view commented that surveys provided information about consumers' opinions, whereas the court was charged with making an assessment of the facts of a situation. Given this, and the fact that surveys documenting perceptions were vulnerable to criticism, these respondents felt that the reduced reliance on survey evidence was easy to understand and unlikely to change.

458. Some also noted that the increasing costs of litigation, and the increasingly cluttered court system, meant that cases were often resolved at lower levels, using alternative dispute resolution procedures. In addition, while they considered that the introduction of the Fair Trading Act 1986 had led to an initial flurry of cases, some of which adduced survey evidence, knowledge of the Act had increased, the "rules" had become better known, and the number of cases taken and heard at higher levels had decreased as a consequence.

459. A small number of lawyers were very experienced users of survey evidence and they noted an increasing reliance on pilot surveys to "test the water" and assess whether preliminary results supported the development of a more extensive survey. This approach recognised both the cost and the risk of commissioning surveys, and the

knowledge that unhelpful survey results can rebound and complicate a case. However, although pilot surveys were considered a useful guide, most lawyers felt it unwise to adduce evidence based on a pilot, since the sample sizes were typically small (often fewer than 100 respondents).

460. When deciding whether to commission a survey, respondents considered whether opposing counsel had conducted a survey that produced findings they felt were open to interpretation and dispute. In situations such as these, a more thoughtful and rigorous survey could not only demonstrate the limitations of other evidence but could strengthen the arguments they were advancing.

461. Lawyers also considered whether the mark in dispute was a significant one and whether the survey findings could be useful in other jurisdictions. If these criteria were met, they were more likely to commission a survey; however, if the mark or dispute was likely to be confined to New Zealand, the overall size of the market often meant that the cost of commissioning a survey could not be justified.

462. The role lawyers played in a dispute also influenced their willingness to adduce survey evidence. Where the onus was on them to demonstrate the distinctiveness of a mark, or the confusion likely to arise from a claim, they thought they would be more likely to find survey evidence a useful addition to other evidence they would use. Where they were defending these claims, lawyers generally felt they would be less likely to commission survey evidence, particularly if they believed the evidence submitted by opposing counsel had obvious flaws that they could expose.

463. Many lawyers commented that even where survey evidence was considered helpful, they could not rely on its merits being obvious to a judge. To ensure the key conclusions from the survey were accepted, lawyers noted that they needed to retain experts who could provide independent and informed comment on the rigour of the survey. Because expert witnesses' role was to assist the court, their comments were typically seen as more independent and carried greater weight as a consequence.

464. For some, survey estimates had little inherent value, although the survey process was considered a valuable means of recruiting consumers who could then appear as witnesses where they could expand on their experience with a particular brand. In confusion cases, lawyers suggested the best evidence was actual evidence of confusion and a witness who had been affected by a claim was likely to be more compelling than a survey, even though the latter may summarise the views of several affected consumers. To the group holding this view, tangible evidence of confusion that could be directly tested was seen

as more powerful than evidence from a wider group if the views and experiences from members of that group could not be individually perused.

465. Nevertheless, while some respondents placed a high value on presenting actual confused consumers to a court, they also recognised that this process was not risk-free, that confused consumers may be unsafe witnesses, and that the courtroom environment could be considered so far removed from the situation in which confusion had allegedly occurred, that the evidence given lacked validity.

466. However, where the question under consideration involved assessment of a mark or attribute, survey evidence was considered more helpful as it could answer questions such as what a mark meant to consumers. Some respondents felt that general open questions such as "what does this mark mean to you?" would be credible and that views from a representative cross-section of consumers could provide insights into the key question judges and hearing officers must address.

467. Yet while consumers' views were considered important, some respondents also noted the importance of other variables. For example, they felt that the value of the product and the type of retail or distribution outlets used were also important, since these provided an indication of the harm that could be caused by confusion and the proportion of a population likely to be at risk of confusion.

468. Nevertheless, consumer surveys could also add information to the wider array of issues that judges reviewed and some respondents thought that consumer evidence could reduce the risk that judges would rely on their own "gut feel", which they thought may not necessarily reflect consumers' views or likely behaviours. To achieve this benefit, they considered that surveys could be most useful if they employed scenarios where consumers were exposed to what one respondent described as the "whole factual matrix", and where an artificial emphasis was not placed on the disputed attribute.

469. However, despite the benefits that could accrue to surveys, some respondents saw the cost of commissioning these as near-prohibitive, particularly when the need for academic experts was included in the overall budget. The cost of conducting a survey with an adequate sample size, that included appropriate and rigorous quality assurance procedures, and that could withstand detailed critical scrutiny was sometimes seen as not worth the risk that the findings could prove unhelpful. This tension between the cost of a survey and the unpredictability of the likely benefits led some respondents to adopt a risk-averse approach and to avoid surveys unless they felt that consumer evidence was critical to a case.

470. Although lawyers recognised the risk that survey evidence may not support the arguments they wished to advance, they nevertheless recognised that some of the "rules" relating to surveys had become more clearly documented over the last decade. In particular, they noted the availability of checklists in some decisions, and among other things, noted the recognised need to identify the relevant universe, avoid leading questions, ensure the questions asked were appropriate, and verify that the interviewers had been correctly briefed and trained, and had not been apprised of the purpose of the survey.

471. Where a decision to proceed with a survey was taken, lawyers' involvement in the survey design process varied. While some wrote and designed in-house surveys, others closely oversaw the work conducted by research companies and expert advisors, while yet others devolved the process and relied heavily on the expertise of those more familiar with survey research. Some lawyers noted that past experience suggested they needed to liaise very closely with the research team to ensure the brief was followed appropriately. However, others felt it was necessary to retain some distance to ensure they were not open to allegations of influencing the research process.

472. As well as commenting on the general advantages and disadvantages that they associated with survey evidence, respondents were also asked their opinion of the different types of error affecting survey evidence and the extent to which they thought these errors could reduce the value of a survey.

6.4.2 Coverage Error

473. Respondents recognised that surveys needed to include both respondents who might currently be purchasing a dispute brand or other brands within the overall product category and respondents who might be future users of the category. However, they felt that the relevant population depended on the issue in dispute and that each case needed to be assessed individually. Almost invariably, therefore, respondents noted that the sample needed to be tailored to the particular dispute being heard. Apart from ensuring the population included likely (or interested) as well as current users of a product category or particular brands within this, respondents did not employ any rules when defining the survey population.

474. Like the other groups interviewed, lawyers did not have a particular rule of thumb for defining the relevant population. As noted, they believed this should include both actual and potential purchasers, although they considered that the population would depend on the nature of the dispute and the particular product or service involved. In particular, they believed this question required an assessment of the kinds of consumers active in the

product category and purchasing the brand, the market structure, and the level of competition that prevailed.

475. Most commented on the desirability of nationwide samples, so that consumers from different geographic areas, and different supply contexts, were included in the sample. However, the details of which areas were chosen and the relative weighting given to different locations was left to the research company and academic expert, if the latter had been retained.

476. Respondents generally accepted that screening questions could be an appropriate means of defining the relevant sample members from a wider population, although those familiar with screening questions felt their use would depend on the product category in dispute and the purpose of the survey itself. However, if a mall-intercept survey was considered the most efficient means of recruiting a sample for a product category or brand that had limited penetration, screening questions were seen as logical and appropriate. Similarly, if the population had tightly defined demographic characteristics, screening questions were seen as a useful way of ensuring that only eligible respondents were interviewed.

477. Although respondents who had used surveys including screening questions were familiar with arguments for the use of these questions, they were also aware that screening questions could be challenged, particularly if the two sides to a dispute had developed different definitions of the relevant population. For this reason, respondents noted that it was important to have a clear and compelling explanation to support the use of screening questions. In particular, they noted the danger of defining a relevant population so tightly that the sample included only consumers with quite specialised knowledge of the product category or brand in dispute, and who were arguably no longer representative of consumers in general.

478. Some also noted the importance of considering members of distribution channels, particularly where retailers may be responsible for advising consumers. In cases where consumers' decision was likely to be informed by retailers' advice, these respondents noted that any confusion among retailers would be likely to be transferred to consumers. In these situations, information about competing products and the extent of channel overlap could help determine whether surveys with retailers and distributors would be necessary.

6.4.3 Sampling Error

479. Many respondents were unsure what sampling error meant or how it was calculated. Some expressed scepticism that a sample of 300 individuals could adequately represent

the views or behaviours of a much wider group and seemed unaware of the basic principles of sampling. For many, the larger the sample, the more acceptable they thought the survey would be. When asked about the relationship between the sample size and the survey population and whether smaller samples might be more acceptable when the population of interest was small, respondents still considered that samples of fewer than 500 individuals were likely to be open to question. For reasons they were often unable to explain in detail, many respondents favoured a sample size of 500 as a general "rule of thumb". However, despite relying on this general benchmark, most respondents also indicated that they were guided by experts they retained and by market researchers, whom they believed were better placed to determine the relevant sample size and ensure that the resulting error margins would be acceptable.

480. While not all respondents made comments like these, there was a lack of detailed understanding of sampling and error margins among the lawyers interviewed. Some respondents were suspicious of small estimates believing (incorrectly) that estimates of less than 4% could fall "within the margin of error". In fact, each survey estimate has its own error margin and it is theoretically and logically impossible for an estimate to "fall below the margin of error".

481. After clarifying the nature of error margins and how these apply to each survey estimate, respondents commented on what they saw as being acceptable error margins. For many, the commonly quoted 3% error margins that are associated with political opinion polls of 1000 people appeared to set a norm. However, to achieve a maximum error margin of only 3%, researchers would need to ensure the sample size was at least 1000, and larger if less efficient sampling procedures such as clustering had been employed. The cost of interviewing such large samples would be very high, and would exacerbate the risk of commissioning survey evidence.

482. Others commented that surveys with error margins greater than 5% ran the risk of producing indeterminate results where the margin between the estimates was low. To these respondents, the need to ensure that potentially contestable results could be clearly separated was very important, thus low error margins provided them with evidence they felt could be critical during the case.

483. Other respondents felt that the sample ought to reflect the number of purchasers within a product category. Thus, where a product category had high penetration and was purchased by a large proportion of people, the size of the sample should be correspondingly larger. Again, this assumption is at odds with the sampling principles, which state that even small samples can be representative of a much larger population

assuming sample members have been selected in such a way that each individual has a known and non-zero chance of being selected.

484. For some respondents, the overall size of the relevant population was also a factor that they considered when identifying the relevant sample size and some commented that, where the relevant market was very small (such as specialist suppliers or producers), the sample size should not be expected to be large. Some respondents felt that the sample should thus comprise a "significant number of people" but that the definition of what constituted a "significant number" varied according to the nature of the market.

485. Overall, although respondents recognised the importance of sampling error, their overall understanding of this concept and its practical import was weak. Questions of which sampling procedure would be used were typically left to experts and research companies, and few understood the implications of different sampling methods for error margins. Many recognised that their expertise did not extend to this topic and noted that they relied upon market researchers and academic experts to ensure that the samples were appropriately drawn. However, others commented on the need for greater knowledge of survey research techniques among their profession and suggested that more education would assist them and judges to make informed and clear decisions about the value of survey evidence.

486. Respondents did not have strong views about the different survey modes that could be used, although most indicated that the mode that best approximated an actual purchase context would be preferable. Some noted that face to face surveys conducted in a respondents' own home seemed quite removed from the retail context where they would decide whether and what to purchase. However, while they noted that an ideal situation might see retailers agree to allow interviews to be conducted in their premises, this could be difficult to arrange, and thus less likely to be used, particularly if the evidence was required for an urgent hearing.

6.4.4 Non-Response Error

487. Respondents were much less concerned about response rates than they were about sampling error and few noted that they reviewed survey response rates in any detail; again, the determination of an acceptable response rate was left either to experts to assess, or research companies to determine. To those with this view, the overall sample size rather than the number of contacts required to achieve a particular sample, was the critical issue.

488. While most respondents did not have a firm view about acceptable response rates, they agreed that the higher the response rate the better. Those with more experience of using survey evidence were more willing to suggest unacceptable response rates and felt that any survey that failed to achieve at least a 50% response rate would be easily criticised. Despite this, respondents commented that a range of factors could affect a survey's response rate and suggested the response rate would depend on the type of product or the market under investigation and the perceived sensitivity of the questions asked.

489. However, some also commented that because survey response rates were declining, and because interviewers often telephoned at inconvenient times, judges themselves may be unwilling to participate in surveys and may view those who do have the time, energy and inclination to respond as atypical consumers. If this is the case, the value of survey evidence could be called into doubt. Overall, however, respondents felt that the courts did not use benchmarks to assess response rates and that judges' view of what an acceptable response rate was would be largely determined by the arguments presented.

490. Although few routinely expected to see an analysis of non-response error in survey findings, several respondents agreed that such an analysis would be useful and could assist judges to understand the contribution a survey may make to a case. However, they also noted that an overly technical analysis of non-response error could be unhelpful and that consideration of this issue would need to be part of a larger argument put forward about the merits (or limitations) of a survey.

6.4.5 Measurement Error –Questionnaire Design

491. Although most lawyers made an initial assessment of whether consumer survey research could assist the case they were developing, most contacted a research company or an academic expert and briefed them on the case. At this stage, lawyers gave survey research experts responsibility for developing the questionnaire, although they also ensured that the questions were appropriately designed and could address the legal question of interest.

492. Although lawyers were very concerned to avoid leading questions, they tended to rely on market researchers and academics to ensure that the survey questions were properly designed and the questionnaire logically and appropriately structured. Several noted that the questionnaire had to have a clear and compelling logic and that the researchers and experts needed to provide a strong rationale to support both the nature of the questions asked and the order in which these were presented. Respondents expected the research team (including members of the market research company responsible for collecting the data and experts charged with overseeing and advising on this process) to present them

with a clear brief in which they explained how the sample would be selected and how the integrity of the results would be ensured.

493. While concerned to avoid leading questions, some lawyers were confident that judges were open to being persuaded by survey findings and that the task of the overall legal team was to ensure that the findings presented were reliable, valid and compelling. Respondents preferred using open-ended questions, and favoured broadly expressed questions at the beginning of a survey, since these had often been accepted as non-leading in earlier decisions.

494. However, lawyers felt that judges might be more likely to accept the use of closed-questions in large surveys, where the task of coding several open-ended questions could become quite onerous. Yet, while respondents accepted there could be a place for closed-questions, some noted that any question that could be answered "yes" or "no" would be considered leading. To these respondents, closed questions should provide respondents with a list of options, rather than a single option with which respondents could agree or disagree.

495. Lawyers also accepted the need to create a context for survey questions and did not necessarily see this as problematic or likely to lead to allegations that the question had become leading. However, like other groups interviewed, they were wary of developing a context that was so specific it directly suggested an association between a brand and a particular attribute, or a specific interpretation of a claim, particularly if this context was created early in the survey. Some noted the tension between ensuring questions were non-leading on the one hand, but providing sufficient context to ensure respondents' answers were not meaningless. However, they noted that judges had recognised the need for respondents to be taken to the issue at the heart of a dispute, although they remained cautious about the extent to which researchers could direct respondents without leading them.

496. For some respondents, avoiding the creation of the very association the survey was designed to investigate depended on the product category in question. This group suggested that since a large brand could be considered synonymous with a product category, drawing respondents' attention to the category could effectively highlight the main brand within this. The same problem was less likely to occur in product categories that were not dominated by a small number of brands, although even in these categories, the level of specificity provided in questions could risk highlighting particular brands and thus leading respondents to a specific answer.

497. Lawyers believed that the survey should be as simple and pared-back as possible and several noted that collection of details not directly related to the legal issue could ultimately prove unhelpful. Like some of the academic experts, they believed that questions involving complex analytical procedures could be counter-productive and that any additional insights the results might provide could be lost in the quagmire of detail they and judges or hearing officers would have to comprehend.

498. Several respondents noted the importance of ensuring the questions addressed relevant legal concepts. For example, some noted that trademark cases would not necessarily be supported by evidence of brand and attribute association, since this was only the first step in testing whether an attribute functioned as a badge of origin.

6.4.6 Measurement Error – Interviewer Quality Assurance

499. Respondents recognised that interviewers played a crucial role in ensuring the quality and integrity of the data and they noted that interviewers should not lead respondents in any way. Many commented on the fact that a questionnaire could be non-leading, but that interviewers could, through comments and gestures, lead respondents. In particular, they felt it was very important that respondents were not co-erced in any way, either to take part in the survey or to provide an answer that the interviewer had suggested would be more suitable than other options. For this reason, they felt that interviewer training was critical and most supported higher levels of quality assurance than would normally be applied to surveys.

500. Quality assurance procedures such as providing video or audio evidence of training provided prior to the data collection and details of each interviewer's experience and training record were considered helpful ways of establishing the credibility of the survey. Details of site visits undertaken and observations of the interviewers' conduct were also thought likely to support the integrity of the data, although most devolved decisions about the extent of interviewer supervision in the field to either the research company or the expert who would be commenting on the survey findings.

501. Although lawyers agreed that full details of the survey procedure as well as the actual findings had to be available for scrutiny, not all believed that video records of interviewing or training would be required. However, where the interviews were to take place in several different centres and the project director was not able to brief each set of interviewers in person, respondents agreed a training video that ensured a systematic approach to interviewer training and briefing would be helpful.

502. One respondent suggested that, if money was not a limiting factor, tape recorded interviews or video evidence of successful interviews could be helpful, but indicated this

was a "Rolls Royce" approach that judges and hearing officers did not routinely expect. Others were less sanguine about what the video could expose, and suggested that this type of evidence could prove more problematic than beneficial. Nevertheless, all agreed that details of the brief provided to interviewers, the field instructions they received, and the audit procedure used to verify the interviews, should be provided.

503. Yet, while lawyers recognised the importance of documenting each stage of the survey process, several commented that research companies had not always kept appropriate records and that the integrity of some evidence had been damaged because of poor record-keeping. Respondents who had experienced these problems were typically more likely to suggest close oversight of the entire research process was necessary.

504. However, where smaller samples were used and open-ended questions employed, they felt judges would wish to see copies of the actual questionnaires as well as interview transcripts to satisfy themselves that an appropriate quality of interviewing had been maintained. In addition, a visual analysis of the coding frames and questionnaires would enable a cursory assessment of the extent to which the coding had captured variation in responses.

505. Others felt that if a heavy reliance was to be placed on responses to open-ended questions, the researchers should provide audio transcripts of these, so that the judge or hearing officer as well as opposing counsel could check the questionnaire records against the actual interview. Some noted that, if their role was to evaluate a survey presented by opposing counsel, they would see the absence of audio tapes as a serious omission in the material disclosed.

506. These respondents noted that audio recordings, particularly of open-ended questions, would enable them to assess respondents' tone of voice, how consistent this was with their answers, and the level of certainty they had in their responses. Perhaps more importantly, these lawyers suggested that tape recordings would document any prompts given by interviewers as well as any other deviations from the questionnaire script.

6.4.7 Levels of Confusion and Association

507. When asked about the level of confusion or association required to support passing-off claims or trademark applications, lawyers found it difficult to specify precise levels and noted that each case had to be assessed on its merits. In general, the level of confusion that had to be established was lower than the level of distinctiveness required to demonstrate that an attribute functioned as a badge of origin.

508. Of those prepared to comment on benchmark levels, lawyers considered the required threshold was lower in cases of consumer confusion. Here, they suggested estimates that fell below 10% were unlikely to be considered material, unless the likely harm that misled consumers could suffer was severe. However, this figure was not a firm benchmark and they indicated that the actual level considered appropriate by a judge would depend also on the type of product or service, and the relevant consumer group, particularly their knowledge of the product category and vulnerability.

509. Several noted the need to demonstrate that a "substantial number of people" had been misled, or were at risk of being misled, but found it difficult to suggest what level of confusion would be required to meet this threshold. Some considered that at least 15% of a sample should be confused for confusion to be established.

510. Where trademark applications were being considered, lawyers considered that a much higher benchmark should be met before a mark should be considered to be functioning as a badge of origin. Some suggested that a minimum 50% level of association should be established, and that, where the attribute in question had an inherently low level of distinctiveness, such as a colour, the benchmark level should be considerably higher than 50%, although the actual level would depend on the category and attribute in question. When asked why they set this level, respondents indicated that it reflected a clear majority of the sample, although this figure did not appear to be derived from recent decisions.

6.4.8 Role of Experts

511. Experts can fulfil at least two roles in intellectual property disputes. First, they may provide access to more advanced research methodologies than those typically employed by research companies. For example, they may have access to specialised software and the knowledge of sophisticated modelling approaches that estimate behaviour in different contexts. Respondents were asked to comment on the extent to which this knowledge might extend the current use of survey evidence.

512. Because several respondents had commented on the desirability of simple and direct surveys, they were less enthusiastic about the use of more advanced statistical modelling in surveys. Although some recognised that procedures such as stated preference choice experiments could provide greater behavioural insights, they noted that those insights would need to be clear and accessible, and that disputes over the methodology employed could lead to a highly technical discussion that judges and legal counsel may not have the background to appreciate. As a general rule, they felt that research methodology needed to be "intuitively obvious" so that intelligent lay people were easily able to understand the decisions taken and how these contributed to the outcomes presented. For this

reason, most respondents took a very conservative approach when considering survey questions and methodologies, and tended to use designs and questions that had been accepted in past decisions.

513. Experts could also provide direct assistance to a judge or hearing officer and help interpret survey evidence and any more advanced analyses that might be adduced. When asked whether they would support the appointment of a survey expert whose role was to assist the judge, most lawyers were ambivalent. Although they recognised that both sides currently retained their own experts, and that it could be difficult to ensure experts were not bound by the interests of the sides they were representing, they nevertheless had reservations about court appointed experts.

514. However, they acknowledged that advice from court-retained experts could be useful, particularly on technical matters and in establishing the facts that a judge or hearing officer would then use to determine the case. However, some were unsure about whether the advice could be truly objective and suggested that even experts would bring their own frame of reference to the task. That is, they questioned whether true independence or objectivity could be brought to bear on cases and preferred the status quo, which ensured any vested interests could be openly tested.

515. Perhaps more importantly, lawyers questioned the relinquishment of control that they believed would inevitably occur if experts' role expanded to include an advisory function. Litigators, in particular, considered that a court appointed expert could impede the ability of their clients to establish and support their case. While they recognised the value in judges having access to advice that was not associated with either of the parties to a dispute, they felt that this advice should be provided in a public forum so that it could be challenged and tested. They noted that, if an expert sat on the bench with a judge, the expert's own philosophy would not be exposed and open to scrutiny and some thought this was unsafe when compared to the current system, which allows such scrutiny to occur.

516. However, others noted that *amicus curiae* were already used in other types of cases and that this approach would not necessarily be novel. Others commented that a "hot tub" approach had been used in other cases where the parties involved had adduced evidence from several experts whose opinions had varied. A "hot tub" involved bringing all experts together and enabling them to be questioned collectively, rather than in sequence, so that issues in dispute could be explored in detail with the individuals offering different opinions. However, while respondents familiar with this process thought it could bring many benefits, some suggested there could be value in having an independent third party with expertise in the topic chair the discussion.

517. When asked about the prospect of jointly commissioned surveys, several lawyers noted that this approach was at odds with the adversarial system of justice. Others felt that while the idea was appealing in principle, agreements might subsequently be challenged and could form the basis of an appeal. This concern was noted by several respondents, some of whom suggested that there could be a need for a contract between the two parties to ensure that the survey design that was agreed upon did not become disputed should the results prove less helpful to one side than was anticipated.

518. Alongside this concern, some lawyers commented that they would not want to be in a position where they had agreed to a methodology that could later compromise their client's interests. This group suggested that any process of agreement would involve compromises, and these could result in a design that, from their client's point of view was sub-optimal. For these respondents, the risk that they might agree to a design that prejudiced their client's interests was seen as more problematic than the risk that they could commission a survey that was subsequently challenged and accorded little weight.

519. Some respondents noted situations in which the side preparing a survey would adduce details of the proposed sampling procedure and questionnaire in advance. The opposing party to the dispute would then have an opportunity to evaluate and critically review the details provided. If comments made were thought to have substance, the survey could be reviewed to address the criticisms, although the party undertaking the survey would not be under any compulsion to address the criticisms. For respondents familiar with this approach, it was preferable to an agreed upon survey design, since the design would not have to be approved by both parties, although respondents noted that failure to address criticisms that a judge subsequently considered relevant could reduce the weight given to the survey.

6.5 Conclusions

520. Despite having clear views about some issues, and despite their own experience, which sometimes involved several cases in which consumer survey evidence had been adduced, most respondents made it clear that they relied heavily on academic experts to guide the survey design and evaluation processes. While they might oversee the process, and have a detailed involvement in some stages, they nevertheless were guided by experts who they retained to design and execute the survey, interpret the results, and evaluate arguments and evidence produced by opposing counsel.

521. Overall, lawyers recognised the benefits that compelling survey evidence could bring, but many felt that the costs made the use of consumer research a high risk decision. As a result, while none would rule out the possibility that they might commission a consumer

survey to assist the development of a case, many preferred to rely on other evidence, such as details of the market structure, evidence of sales patterns over time, and details of market participants' behaviour (such as their use of advertising and promotion campaigns). These latter details are typically available from secondary data sources and so are less expensive to obtain than primary data collected for the sole purpose of the dispute. Furthermore, because secondary data documents actual behaviour or investment, it is less easily disputed. Nevertheless, where direct consumer evidence was required and existing secondary sources could not inform the dispute adequately, respondents generally agreed that this should be developed from a thoughtfully designed and rigorously implemented survey.

7. Conclusions

522. The principles of survey research apply equally to forensic research and routine market research, the only difference is that forensic studies attract closer, and often very critical, scrutiny and researchers are more accountable for the decisions they take. Analysis of academic literature as well as cases from both New Zealand and international fora suggest that the criticisms levelled at survey evidence have surfaced in many different jurisdictions.

523. The criteria outlined in Appendix 3 aim to assist researchers and lawyers to gain a better understanding of each other's craft, enabling researchers to address the legal question of interest and lawyers to understand the limitations of survey evidence. We hope that the criteria will assist lawyers and their research teams to address the criticisms that have affected the weight given to surveys adduced in New Zealand intellectual property proceedings and to manage these more effectively.

References

Allied Liquor Merchants Lit v Independent Liquor (NZ) Ltd (1989) 3 TCLR 328.

American Association of Public Opinion Research. http://www.aapor.org/AAPOR_council_guidelines.pdf, Accessed 20 June, 2004.

Anheuser Busch v Budejovicky Budvar National Corporation [2001] 3 NZLR 666.

*ARA v Mutual Rental Cars (Auckland Airport) Ltd (*1987) 2 TCLR 141.

ASB Bank Ltd v Trust Bank Auckland Ltd (1989) 3 TCLR 70.

Austin, Nichols & Co Inc v Stichting Lodestar, CIV 2004- 485-1281.

Automobile Club de L'Ouest, Aco v South Pacific Tyres New Zealand Limited CIV 2005 485 248.

Barksdale, H. (1959). The use of marketing data in courts of law. *Journal of Marketing*, 23, (April), 376- 385.

Belson, W.A. (1981). *The design and understanding of survey questions.* Aldershot, England: Gower.

Blenhaven; International Cellars (Marlborough) Ltd v Montana Wines Ltd (1989) 3 TCLR 115.

Bluebird Foods Ltd v Cerebos Greggs Ltd (1998) (CP323/98).

Bottomley, D. (2001). A right-facing crocodile versus a left-facing crocodile: A survey technique for the high court of Hong Kong. Paper presented at the World Association of Public Opinion Research Conference, Rome, 20-22.

Brient, V. and Hebert, W. (2000). Fair use defense applies in aftermarket situations. *National Law Journal*, NLP IP Company.

Carter Holt Harvey Ltd v Cottonsoft Ltd [2004] 8 NZBLC 101

Caughey, R. (1956). The Use of public polls, surveys and sampling as evidence in litigation, and particularly trademark and unfair competition cases. *California Law Review*, 44 (3), 539-547.

Cerebos Greggs Ltd v Unilever New Zealand Ltd (1994) 5 NZBLC 103.

CIBA-GEIGY v Douglas Pharmaceuticals Ltd (1986) 2 TCLR 346

Cohen, D. (1991). Trademark strategy revisited. *Journal of Marketing*, 55, 46-59.

Comité Interprofessionel du Vin de Champagne v Wineworths Group Ltd (1991) 2 NZLR 432

Commerce Commission v Griffins Foods [1997] DCR 797

Cookie Time Ltd v Griffins Foods Ltd [2000] M1756/SW00

Crespi, I. (1987). Surveys as legal evidence. *Public Opinion Quarterly*, 51, 84-91.

Customglass Boats Ltd v Salthouse Bros Ltd [1976] 1 NZLR 36

Davies I. (1995a). Legal Update. *The Journal of Brand Management*, 3(August), 65- 72.

Davies I. (1995b). Legal Update. *The Journal of Brand Management*, 3(October), 121-128.

Deeth, D. (2001). Survey evidence in Canada. Presentation to FICPI Open Forum held in Rome, November 14-17, 2001. Downloaded 15 April, 2003 www.ficpi.org/library/APAA_FICPI_Newport/T5_Deeth.ppt

Dillman, D. (2000). *Mail and Internet Surveys: The Tailored Design Method*. 2nd.ed. New York: Wiley.

Gastwirth, J. (2003). Issues arising in using samples as evidence in trademark cases. *Journal of Econometrics*, 113, 69-82.

Granny May's Management Pty Ltd v Whitcoulls Group Ltd (1992) 5 TCLR 148.

Eko, L. (1998). Trademark parody and the mass media: Going beyond survey evidence in the determination of "a likelihood of confusion". *Communication Law and Policy,* 3, 589-608.

Federal Judicial Law Centre (2004). The Manual for Complex Litigation. Washington DC: US Government Printing Office.

Folsom, R. and Teply, L. (1988). Surveying "genericness" in trademark litigation. *Trademark Reporter*, 78, 1- 31.

Ford, G. (2005). The impact of the *Daubert* decision on survey research used in litigation. *Journal of Public Policy and Marketing*, 24(2), 234-252.

Foxman E., Muehling D. and Berger P. (1990). An investigation of factors contributing to consumer brand confusion, *Journal of Consumer Affairs*, 24 (1), 170-189.

Foxman, E., Berger, L. and Cote, J. (1992). Consumer brand confusion: A conceptual framework. *Psychology and Marketing*, 9 (2), 123-141.

Friskies Ltd v Heinz-Watties Ltd 2 NZLR 663.

Grenier, F. Evidence in trade mark cases. http://www.robic.com/publications/Pdf/213-FMG.pdf#search=%22%22Cordon%20Bleu%22%20%22Bradley%22%20expert%20survey%22

Halford-Harrison, R. and Perkins, P. (2004). A very different approach: A comparison of the use of surveys in trademark actions in the UK and the US. World Intellectual Property Report.
http://www.sjberwin.com/location/london/practicearea/trade_marks/publication/a_very_different_approach_a_comparison_of_the_use_of_surveys_in_trademark_actions_in_the_uk_and_the_us.html

Harris, R. (2002). Surveying the boundaries: Recent developments in trademark surveys. *The Computer and Internet Lawyer*, 19 (5), 17-25.

Hewlett Packard Co. v Xerox Corp., No. C97-3850 SI (N.D.Calif. July 15, 1998).

Hoek, J. and Gendall, P. (2003). David vs Goliath: An analysis of survey evidence in a trademark dispute. *International Journal of Market Research.* 45 (1), 99-121.

Hudis, J. (2000). Experts in intellectual property cases: A new paradigm. www.oblon.com/Pub/display.php?hudisjptosoct00.html

Imax Corporation v Village Roadshow Corporation Limited (CIV. 2005-404-3248).

Imperial Group plc v Philip Morris Ltd [1984] RPC 293

Interlego AG & Anor v Croner Trading Pty Ltd (1991) 102 ALR 379.

International Cellars in Blenhaven; International Cellars (Marlborough) Ltd v Montana Wines Ltd (1989) 3 TCLR 115.

Jacoby, J. and Szybilli, G. (1995). Consumer research in FTC versus Kraft (1991): A case of heads we win, tails you lose? *Journal of Public Policy and Marketing*, 14(1), 1-15.

Jacoby, J. and Hoyer, W. (1990). The miscomprehension of mass-media advertising claims: a re-analysis of benchmark data. *Journal of Advertising Research*, 30 (3), 9-17.

Jacoby, J. and Morin, M. (1998). "Not manufactured or authorized by…": Recent federal cases involving trademark disclaimers. *Journal of Public Policy and Marketing*, 17 (1), 97-107.

Kapferer J-N. (1995b). Brand confusion: Empirical study of a legal concept, *Psychology and Marketing*, 12(6), 551-568.

Kearney, I. and Mitchell, V-W., (2001). Measuring consumer brand confusion to comply with legal guidelines. *International Journal of Market Research*, 43(1), 85-91.

Keller, B. (1992). A survey of survey evidence. *Litigation,* Fall.

Klissers Farmhouse Bakeries Ltd v Harvest Bakeries Ltd (1985) 2 NZLR 143.

Klissers Farmhouse Bakeries Ltd v Harvest Bakeries Ltd [1988] 2 TCLR 555.

Koerner, R., 1980. The design factor – An under-utilised concept? *European Research*, 8(6), 266-272.

Leiser, A. and Schwartz, C. (1983). Techniques for ascertaining whether a term is generic. *Trademark Reporter*, 73, 376-390.

Lessem, J. (2000). Consumer surveys in trademark litigation. http://www.cll.com/articles/article.cfm?articleid=36

Levi Strauss Co. v Kimbyr Investments Ltd (1994) 1 NZLR 332.

Loken, B., Ross I. and Hinkle, R. (1986) Consumer "confusion" of origin and brand similarity perceptions. *Journal of Public Policy and Marketing*, 5, pp. 195-212.

Magellan Corporation Ltd v Magellan Group Ltd (1995) 6 TCLR 598.

Mainland Products Ltd v Bonlac Foods (NZ) Ltd (1998) 8 TCLR 224.

Market Milk Federation of New Zealand Inc v Woolworths (New Zealand) Ltd (1992) 4 TCLR 619.

Maronick, T., 1991. Copy tests in the FTC deception cases: Guidelines for researchers. Journal of Advertising, 31, 9-17.

McCarthy, T. (1998). *Trademarks and Unfair Competition*, 4th ed. Rochester, NY: Lawyers Co-Operative Publishing Co.

Miaoulis G. and D'Amato N. (1978). Consumer confusion and trademark infringement. *Journal of Marketing*, 42, (April), 48-55.

Morgan, F.W, (1990). Judicial standards for survey research: An update and guidelines. *Journal of Marketing*, 54, 59-70.

Morin, M. and Jacoby, J. (2000). Trademark dilution: Empirical Measures For An Elusive Concept. *Journal of Public Policy and Marketing*, 19 (Fall), 265-276.

Noel Leeming Television Ltd v Noel's Appliance Centre Ltd (1985) 1 TCLR 290.

Patience & Nicholson (NZ) Ltd v Cyclone Hardware Pty Ltd (2001) 3 NZLR 490.

Pflüger, A. (2001). The misled consumer: Empirical legal research in Germany throughout the change in case law of the European Supreme Court. Paper presented at the 2001 WAPOR conference, Rome, September 22.

Pioneer Hi-Bred Corn Company v Hy-Line Chicks Pty Ltd [1975] 2 NZLR 422.

Pitstop Exhaust Ltd v Alan Jones Pit Stop International Ltd (1987) 2 TCLR 502.

Preston, I. (1992). The scandalous record of avoidable errors in expert evidence offered in FTC and Lanham Act deceptiveness cases. *Journal of Public Policy and Marketing*, 11(2), 57-67.

Re *Estheal; Pierre Fabre SA v Estée Lauder Cosmetics* (1989) 3 TCLR 133

Sarel, D. and Marmorstein, H. (2002). Designing confusion surveys for Cyberspace trademark litigation: The admissibility vs. weight debate. *Intellectual Property and Technology Journal*, 14 (9), 12-17.

Senders, J. and Green, M. (1999). "Any fool can see that these two trademarks are different" ERGO/Gero Human Factors Science. http://www.ergogero.com/pages/IPCogSci.html

Simonson, I. (1994). Trademark infringement from the buyer perspective: Conceptual analysis and measured implications. *Journal of Public Policy and Marketing*, 13, 2, pp. 181-99.

Skinnon, J. and McDermott, J. (1998). Market surveys as evidence: Courts still finding fault. *Australian Business Law Review*, 26, 435-449.

Stewart, D. (1995). Deception, materiality, and survey research: Some lessons from *Kraft*. *Journal of Public Policy and Marketing*, 14 (1), 15-28.

Sudman, S. (1995). When experts disagree: Comments on the articles by Jacoby and Syzbillo and Stewart. *Journal of Public Policy and Marketing*, 14 (1), 29-34.

Swann, J. (1980). The validity of dual functioning trademarks: Genericism tested by consumer understanding rather than by consumer use. *Trademark Reporter*, 69, 357-376.

Swann, J. and Palladino, V. (1988). Surveying "genericness": A critique of Folsom and Teply. *Trademark Reporter,* 78, 179-196.

Taylor, C and Walsh, M. (2002). Legal Strategies for protecting brands from genericide: Recent trends in evidence weighted in court cases. *Journal of Public Policy and Marketing*, 21 (1), 160-167.

Universal College of Learning v ACP Computer Solutions Limited and The College of Future Learning New Zealand Limited, CP 9-01, Oct 22, 2003.

Yves St Laurent Parfums v Louden Cosmetics (1997) 39 IPR 11.

IPONZ Cases

1994/05; IP No. 154636, RITZ, 09 March, 1994.

1997/25; IP No. 177534, Gentle Care, 28 July, 1997.

2002/29; IP No. 313001, I-PROFEN & device, 22 August, 2002.

2002/32; IP No. 56330, FELIX, 04 July, 2002.

2002/36; IP No. 224139, P POLLINI stylised SPA, 25 July, 2002.

2002/44; IP No. 312128, 312129, PAINTDIRECT, 27 August, 2002.

2002/51; IP No. 625055, 625056, 625057, ROBOCUP, 3 October, 2002 .

2002/63; IP No. 608292, KELLY BROWN; KELLY BROWN BEER, 9 December, 2002.

2003/29; IP No. 251691, 3-headed shaver shape, 8 August, 2003.

2003/30; IP No. 302315, 302316, the colour green, 23 July, 2003.

2003/32; IP No. 642574 PURPLE, 30 July, 2003.

2003/43; IP No. 295211, the colour orange, 26 November, 2003.

2004/30; IP No. 657218, B STAR, 20 December, 2004.

2005/15; IP No. 675895, VMAX device, 23 May, 2005.

APPENDIX 1: Calculation of Error Margins

The standard error of a survey estimate is calculated thus:

Standard error = $\sqrt{\frac{p(1-p)}{n}}$

The maximum margin of error that would apply to a survey based on a 95% confidence interval is calculated thus:

Margin of error (95%) = 1.96 × $\sqrt{\frac{0.5(1-0.5)}{n}} = \frac{0.98}{\sqrt{n}}$

Including a design effect would normally see the error margins increase in size. The design effect for a clustered sample can range from 1.5 upwards, depending on the size and number of the clusters. The design effect is taken into account thus:

Design effect * Margin of error (95%) = D.E. x 1.96 × $\sqrt{\frac{0.5(1-0.5)}{n}} = \frac{0.98}{\sqrt{n}}$

APPENDIX 2: Calculating Outcome Rates from Final Disposition Distributions

Numerous outcome rates are commonly cited in survey reports and in the research literature. The same names are used to describe fundamentally different rates and different names are sometimes applied to the same rates. As a result, survey researchers are rarely doing things in a comparable manner and frequently are not even speaking the same technical language. As Groves and Lyberg (1988) have noted, "(t)here are so many ways of calculating response rates that comparisons across surveys are fraught with misinterpretations." Among the more common terms utilized are response, cooperation, refusal, and contact.

As defined by CASRO (Frankel, 1983) and other sources (Groves, 1989; Hidiroglou, et al., 1993; Kviz, 1977; Lessler and Kalsbeek, 1992; Massey, 1995), the response rate is the number of complete interviews with reporting units divided by the number of eligible reporting units in the sample. Using the final disposition codes described above, several response rates are described below:

RR = Response rate
COOP= Cooperation rate
REF = Refusal rate
CON = Contact rate
I = Complete interview (1.1)
P = Partial interview (1.2)
R = Refusal and break-off (2.10)
NC = Non-contact (2.20)
O = Other (2.30)
UH = Unknown if household/occupied HU (3.10)
UO = Unknown, other (3.20)
e = Estimated proportion of cases of unknown eligibility that are eligible

Response Rates

$$RR1 = \frac{I}{(I + P) + (R + NC + O) + (UH + UO)}$$

Response Rate 1 (RR1), or the minimum response rate, is the number of complete interviews divided by the number of interviews (complete plus partial) plus the number of non-interviews (refusal and break-off plus non-contacts plus others) plus all cases of unknown eligibility (unknown if housing unit, plus unknown, other).

$$RR2 = \frac{(I + P)}{(I + P) + (R + NC + O) + (UH + UO)}$$

Response Rate 2 (RR2) counts partial interviews as respondents.

$$RR3 = \frac{I}{(I + P) + (R + NC + O) + e(UH + UO)}$$

Response Rate 3 (RR3) estimates what proportion of cases of unknown eligibility is actually eligible. In estimating e, one must be guided by the best available scientific information on what share eligible cases make up among the unknown cases and one must not select a proportion in order to boost the response rate.[1] The basis for the estimate must be explicitly stated and detailed. It may consist of separate estimates (Estimate 1, Estimate 2) for the sub-components of unknowns (3.10 and 3.20) and/or a range of estimators based of differing procedures. In each case, the basis of all estimates must be indicated. [2]

$$RR4 = \frac{(I + P)}{(I + P) + (R + NC + O) + e(UH + UO)}$$

Response Rate 4 (RR4) allocates cases of unknown eligibility as in RR3, but also includes partial interviews as respondents as in RR2.

$$RR5 = \frac{I}{(I + P) + (R + NC + O)}$$

$$RR6 = \frac{(I+ P)}{}$$

[1] For example, different values of e would be appropriate in a survey requiring screening for eligibility (e.g., sampling adults 18-29 years old). Two different e's might be used for confirmed households that refused to complete the screener (for which we need an estimate of the likelihood of one or more household members being 18-29) and units that were never contacted (for which we need an estimate of the proportion that are households and an estimate of those with someone 18-29)

[2] For a summary of the main methods for estimating e in RDD surveys (1) minimum and maximum allocation, 2) proportional allocation, 3) allocation based on disposition codes, 4) survival methods, 5) calculations of number of telephone households, 6) contacting telephone business offices, and 7) continued calling), see Smith, 2003.

(I + P) + (R + NC + O)

Response Rate 5 (RR5) is either a special case of RR3 in that it assumes that e=0 (i.e. that there are no eligible cases among the cases of unknown eligibility) or the rare case in which there are no cases of unknown eligibility. Response Rate 6 (RR6) makes that same assumption and also includes partial interviews as respondents. RR5 and RR6 are only appropriate when it is valid to assume that none of the unknown cases are eligible ones, or when there are no unknown cases. RR6 represents the maximum response rate.

SOURCE: American Association of Public Opinon Research (2006). Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys. Available at: http://www.aapor.org/pdfs/standarddefs_4.pdf

APPENDIX 3: Criteria to Guide Survey Development and Data Collection

The following questions and criteria are based on the analysis of cases, a review of selected international literature, and interviews with individuals who have expertise in various aspects of intellectual property proceedings. No criteria can guarantee that a survey will be immune to criticism or that a judge or hearing officer will find it persuasive. However, we hope that the questions and suggestions outlined below will reduce the likelihood that a survey fails to "pass first base".

Deciding Whether to Conduct a Consumer Survey

Given the cost and expense of conducting consumer surveys, it is prudent to assess the role consumer survey evidence might play, the risk that the data collected may be unhelpful, and the question of whether other data could perform the same function.

- What other evidence is available and able to be adduced? Could the questions the survey would be designed to address be addressed using existing data?
- What is the level of risk if the survey findings are indeterminate or unhelpful? Does the case hinge on consumer survey evidence? How compelling is this likely to be? Would the case fail if the survey evidence was unhelpful?

Identifying a Research Team

A research team should offer relevant expertise, have a sound track record, and be knowledgeable about the both the product or service area and the research methodology required to address the legal question of interest. To assist a research team, an expert consultant may be engaged; this individual is often an academic who has specific expertise in survey research methodology. The following questions are designed to assist in the selection of a research team.

- Does the research company have experience in forensic research? How have other surveys they have designed or overseen fared in court? What criticisms have been levelled at their work in the past and how do they plan to address these problems?
- Is it necessary to retain an academic expert? Would the research company's work benefit from external peer review at all stages of the research development? Would an external expert review allow more confidence to be placed in the survey findings?
- Does the expert have previous experience in this field? How has her or his evidence been treated in the past? Have questions about the expert's qualifications or depth of expertise been raised?

- If the expert has given evidence in previous cases, is the present case likely to give rise to any conflicts? Could previous evidence be used to create a conflict with the current case? Would the expert be asked to develop, use or comment on a research methodology she or he has previously criticised?

Defining the Survey Objectives

Surveys have been criticised for failing to address the legal question of interest. It is thus critical that the legal team and the research team have a common understanding of the survey purpose.

- What is the legal question that the survey needs to address? Are there any precedents that could be drawn on to help define the legal question or inform the survey?
- Defining the overall survey objective should inform the development of the individual survey questions. Each individual question should be checked against the overall objective to ensure it contributes to answering the legal question.
- Avoid general objectives; try to specify the legal question as precisely as possible, and avoid questions that are unrelated to the overall legal issue.

Identifying the Relevant Population

The relevant population may need to be considered generally, at least initially. End-users may be affected by confusion, as may members of the supply chain. If this is the case, it may be necessary to consider different surveys to estimate confusion or deception among the various relevant groups.

- Who is likely to be affected by any confusion? Consider how different affected parties may be defined and identify the populations of interest that may exist.
- Ensure the groups defined consider both those who may currently be affected by any confusion and those who may be at risk of being affected by deception some time in the future.
- Consider consumers' current behaviour and consider whether the sample is large enough to enable identification of different consumer groups and separate analysis of their responses, should this be appropriate.
- If screening questions are to be used, these should be designed so they avoid alerting respondents to the topic of interest. This may be achieved by including the screening criteria within a list so respondents are unaware which item in the list is the one used to determine their eligibility.
- How dispersed is the population of interest? Recent judgments suggest that regional samples are appropriate; however, if behaviours differ across regions, a nationwide

survey should be undertaken to ensure the affected population is appropriately represented in the sample.

- Consider the most effective way of reaching the populations of interest and ensure that a diverse group of relevant respondents can be identified and interviewed.

Managing Sampling Error

Sampling error occurs because it is not practical to interview an entire population. It may be decreased through increasing the sample size, but may be increased if particular sampling approaches are used. Judgments provide some guidance about the size of relevant samples.

- Sample sizes should comprise at least 300 individuals; samples based on fewer than 300 individuals have been criticised in the past, will have larger error margins, and so may be more vulnerable to criticism.
- Ideally, samples will be chosen using a random selection procedure. However, in some cases, mall intercept sampling may be appropriate and checks should be made to demonstrate that the sample matches the wider population of interest. Comparison of the sample profile with the mall traffic profile and then with the population profile (if this is known) will enable an assessment of whether the sample matches the wider population of interest.
- If weighting is to be used to adjust the sample profile to match the population profile, the weighting variables should have a clear link to the behaviour of interest.
- Quota sampling based on standard age-gender profiles should not be routinely employed unless the profiles used correspond to the population of interest. Where this is not the case, quotas should be used with caution.
- Ensure that the sampling procedure used and the methods for respondent selection have been carefully documented and that interviewers have adhered to these. Review sampling procedures during interviewer training and ensure details of this training have been recorded.
- Where error margins are calculated, these should acknowledge the sampling procedures used and include the appropriate design factor in the calculation. If statistical significance tests are to be undertaken and submitted as evidence, the sample size(s) must be sufficient to ensure the relevant tests can be performed.
- Statistical significance does not necessarily indicate practical significance; even small differences may appear statistically significant if the sample size is large enough, these may have little practical import. Researchers should ensure appropriate distinctions between practical and statistical significance are maintained.

Managing Non-Response Error

Non-response error can undermine the robustness of survey estimates since, if only a small proportion of those approached to participate in a survey actually do so, they may differ from the group who chose not to take part. Although New Zealand cases have not placed a heavy emphasis on the response rate, internationally, the response rates are considered as part of the overall evaluation of survey evidence. For this reason, attempts should be made to ensure response rates reach at least 50%, the minimum deemed acceptable in the US

- Ensure interviewers are well-trained and experienced. Review training material to ensure that it includes information on converting refusals and gaining compliance from reluctant respondents.
- Review interview briefing material and include information on approaching respondents and securing compliance. Role-play refusal conversion techniques to ensure interviewers are familiar with these.
- When using at-home face-to-face interviews and telephone interviews, employ multiple call-backs to decrease the proportion of respondents who could not be contacted. If possible, ensure fieldwork is conducted over a 7-10 day period of time to maximise opportunities for multiple call-backs.
- In mall-intercept surveys, record brief demographic details of refusals (estimate age and record gender). Compare the refusal profile with the achieved sample.
- Ensure the potential effects of non-response are considered and discussed in reports presented to the courts.
- The response rate should be calculated using a known and accepted formula. Where assumptions are made about the proportion of ineligible respondents present in those unable to be contacted, these assumptions should be clearly stated and their use justified.

Managing Measurement Error

Survey Design

- The survey must address the legal question of interest (see Setting Objectives). Lawyers, researchers and experts should consult carefully to ensure they share a common understanding of the survey's overall purpose. Surveys that fail to address the legal question of interest typically carry little weight.
- Where possible, use a methodology that has previously been accepted in the courts. If the courts favour a particular style of question, departing from this approach may increase the risk that the survey will be criticised and its influence reduced.
- If closed questions are used, these must be tested to ensure they contain a full range of response options. Careful pre-testing can assist with this task.

- Ensure that the task respondents are asked to complete approximates as closely as possible the task they would perform in an actual purchase context. Judges have been critical of surveys that lack external validity.
- Recent cases suggest it is desirable to include a control in surveys. This typically takes the form of a brand or other attribute that is unrelated to the legal question of interest. The level of response to the control provides a benchmark against which responses to the disputed attributes may be assessed. If two surveys are required (one containing the test stimulus and the other a control), respondents must be randomly allocated to test and control conditions and the allocation procedure should be documented.
- The survey questions should be designed to ensure they do not lead respondents directly to a given response. While the courts have accepted that researchers need to create a context in which to locate questions, questions that take respondents directly to the issue of interest are likely to be considered leading. A sequence of questions, moving from general to more specific topics, has been considered acceptable in previous decisions.
- Peer review and pre-testing are quality assurance measures that may help minimise the possibility that the survey will rely on leading questions. Pre-testing should be undertaken with a small sample of respondents to ensure the questions are appropriately worded and understood, any skips in the questionnaire function correctly, and the interviewer instructions are clear and unambiguous. Where the survey design or question wording may be disputed, researchers should oversee the pre-testing and could consider using cognitive pre-tests as a further quality assurance measure. Details of pre-testing should be clearly recorded and included in the material submitted to the court.

Interviewer Error

- Ensure all interviewers are well-trained in general survey research procedures. Keep records of interviewers' previous work so their experience and ability can be documented if required.
- If open-ended questions are used, interviewers must be trained to elicit the full range of answers respondents are able to provide. The interviewer briefing session should include discussion of probing and recording of open-ended responses.
- Conduct a detailed briefing session and video this so it can be provided in evidence. Ensure each question is explained and role-played. If the survey contains skips, it may be prudent to role play the questionnaire with each interviewer to ensure the skips are correctly administered.
- Under no circumstances should the purpose of the survey be disclosed to interviewers. Those conducting the interview should not have any interest in the outcome of the survey. There is some evidence that interviewers who are aware of the survey purpose have not administered questionnaires in a disinterested manner.

- Consideration should be given to audio-taping the interviewers while they administer the surveys. Where well-trained and competent interviewers are employed, this procedure should attest to the quality of the data; however, the reverse also applies.

Survey Administration

- Conduct periodic observations of the interviewing so that the quality of this can be documented in material provided to the court.
- Audit at least 20% of the questionnaires and increase the proportion audited if any discrepancies are detected. Ensure that the audit examines not only the administration of the questionnaire (was an individual interviewed) but administration of particular questions (does the respondent remember being asked X). Ford (2005) notes that a 100% audit is becoming the norm in the United States and researchers should be aware that even a 20% audit may be challenged as insufficient.
- The questionnaires will need to be provided to opposing counsel thus at the end of each interview, respondents should be advised the survey may be used in litigation. They should be asked to provide their name and contact details, but should also be able to keep these details confidential. Where the survey content is not confidential, judges have been less sympathetic to arguments that respondent anonymity is required.

Survey Interpretation and Reporting

- Details of any coding schemes developed and employed should be carefully documented. Where the case relies on interpretation of open-ended questions, the coding schemes should be independently developed by at least two coders. The resulting schemes should be compared, any discrepancies identified, and the final agreed scheme should be scrupulously applied.
- The survey report should consider the extent to which the four types of errors may have affected the estimates obtained. Those responsible for writing the report should note any appropriate qualifications and how these might affect the survey estimates.
- Any statistical tests should be clearly set out and explained. The rationale for the test, the results and their implications should be described in such a way that a lay audience can understand the conclusions drawn.

APPENDIX 4: Interview Protocol

Respondents were first asked to discuss recent cases in which they had been involved and where survey evidence was adduced. The discussion was loosely structured around the four types of error outlined in the report and was linked back to the specific cases respondents mentioned during the preliminary stage of the interview.

Coverage and Sampling Error

- Who was surveyed? How many people were interviewed? How were they selected? What were the error margins? What do you understand by an error margin?
- How acceptable was the sample size? What problems, if any, were identified with the selection procedures (issues of randomisation etc.)?
- What do you believe is an appropriate sample size for a survey of the general population? For more specific groups? Respondents will be shown a series of examples and asked to comment on what they see as a relevant sample size.
- How representative do you think mall-intercept surveys are?
- Respondents will be shown screening questions that can be administered as qualifying questions or as final questions and will be asked to comment on the appropriateness of these.

Non Response Error

- What is an acceptable response rate? How do you treat response rates of more than 50%? Less than 50%?
- In your experience, do research companies present information on non-response error? How do you use this? Would you find comparisons on respondents' and non-respondents' demographic traits helpful or not?
- How should response rates be calculated? Respondents will be shown a series of examples and asked to identify the one they believe is most appropriate, and provide reasons for their choice.

Measurement Error

- Respondents will be shown examples of open and closed questions and asked to comment on how robust they believe these are.
- Respondents will be asked to comment on a range of question wording approaches (these will include direct questions, balanced questions, and questions that include specific "uncertain" response options.
- Respondents will be shown a range of closed question response options and asked to comment on how complete these are and how they think completeness should be ascertained.
- What kind of interviewer training is desirable? What experience should interviewers have? What is the balance between fieldwork quality and fieldwork cost that they think is appropriate?
- What level of field supervision is desirable? What other QA measures are desirable to establish the quality of the data? (Discussion of video tapes of training, sound recordings of interviews etc.)

Future Issues

- At present, surveys can be adduced by both sides to a dispute. These tend to cancel each other out. What do respondents think of the suggestion that surveys should be designed in consultation with each side by a court appointed expert? Advantages and disadvantages of this suggestion?

- Extension of survey evidence to include more sophisticated methodologies? Advantages and disadvantages?

- Exploration of other issues arising during the discussion. Respondents will also be asked to comment on other topics that have been mentioned.

The interview protocol was reviewed at the conclusion of each interview so that issues identified by respondents would be explored with subsequent respondents (comments from respondents were treated anonymously).