Using Daubert-Kumho Gatekeeping to Admit and Exclude Surveys in Lanham Act Advertising and Trademark Cases

Richard J. Leighton

“Did He Really Say That?” Survey Evidence in Deceptive Advertising Litigation

William W. Vodra
Randall K. Miller

Assessing Trademark Significance: Genericness, Secondary Meaning and Surveys

Vincent N. Palladino

Experimental Design and the Selection of Controls in Trademark and Deceptive Advertising Surveys

Jacob Jacoby

Litigation Surveys—Social “Science” as Evidence

Michael Rappeport
LITIGATION SURVEYS—
SOCIAL “SCIENCE” AS EVIDENCE

By Michael Rappeport

I. INTRODUCTION

Surveys are often a necessity in Lanham Act litigation. They provide evidence of the perceptions of a cross-section of an appropriate population and, where there is no other evidence of actual confusion, may be the only way to proffer such evidence. However, surveys are inherently imperfect measures of human perceptions. Therefore, they constitute a means of illuminating the likely situation in the real world rather than proof in the usual physical science sense of that term. Surveys are inherently case specific; that is, each survey must be designed in light of the circumstances and theories of the specific case. Because they are inherently imperfect and case specific instruments, the design and implementation of surveys cannot be governed by some set of “forms,” but instead must be guided by the application of general principles to each specific case. For all of these reasons, the value of a survey to a court is rarely a yes or no proposition, but instead a survey must be judged in terms of how well it clarifies the factual issues in a specific case.

In order to judge the quality of a survey, we must develop criteria to use in deciding upon and selecting an appropriate sample of persons to interview, designing and structuring an appropriate approach and questionnaire, conducting the survey, and analyzing and evaluating the results to address the appropriate litigation issues. This article will explore, from the point of view of a survey researcher, how such criteria for the design and conduct of surveys of all kinds can be applied to Lanham Act litigation surveys.

II. THE NEED FOR SURVEYS

Over fifty years ago, Judge Jerome Frank, in a dissent, articulated what has subsequently become the classic judicial argument for why surveys are needed in Lanham Act litigation:

---

* Partner in RL Associates, Princeton, New Jersey, Associate Member of the International Trademark Association; Member of The Trademark Reporter Committee. The author would like to thank Theodore Max, Partner at Mintz Levin Cohn Ferris Glovsky and Popeo PC and Allison Weiss, Summer Associate at Mintz Levin Cohn Ferris Glovsky and Popeo PC, for their help in editing the article.
As neither the trial judge nor any member of this court is (or resembles) a teen-age girl or the mother or sister of such a girl, our judicial notice apparatus will not work well unless we feed it with information directly obtained from “teen-agers” or from their female relatives accustomed to shop for them.\footnote{Triangle Publications, Inc. v. Rohrlich, 167 F.2d 969, 976, 77 U.S.P.Q. 196, 201-02 (2d Cir. 1948) (Frank, J., dissenting). Frank’s dissenting view has become the majority view. “The need for a more certain test of probable consumer reaction than mere judicial hunches was noted as early as 1930 by legal scholars. . . . Thirty years later, the Second Circuit, after noting Judge Frank’s lament, remarked that, ‘The law of unfair trade practices has come a long way’ since then. . . . By 1983, the Fifth Circuit could state that: ‘Survey evidence is often critically important in the field of trademark law. We heartily embrace its use . . . .” 5 J. Thomas McCarthy, McCarthy on Trademarks and Unfair Competition, § 32:194 and nn. 5-6 (4th ed. 2002).}

While it is certainly true that Judge Frank was not a member of the appropriate public, his comment still addresses only one part of the reason why surveys are a necessity in Lanham Act cases. To understand what really motivates and justifies the use of surveys in Lanham Act litigation, consider the following hypothetical. Assume that the case before Judge Frank had been about purchasing judicial robes, and assume that Judge Frank was in the habit of purchasing his own robes. Then, to use his own words, he surely would have been able to feed the judicial apparatus with (appropriate) information, in which case, the problem he discussed in Triangle Publications would disappear, but, nevertheless, the reason to conduct a survey would still exist.

The need for surveys stems from the underlying justification for the granting of the monopoly we call a trademark. Society grants the trademark owner a monopoly in order to make it feasible for consumers \textit{as a whole} to use the trademark as a source and quality identifier for a product or service.\footnote{1 McCarthy, id. at § 3:10 (“the source function remains secure for a very large number of marks where customers perceive the mark as signifying that the goods are supplied by that source that owns the mark.”).} That is, the Lanham Act does not look to whether one individual (even if that individual is a judge) is confused or misled, but rather to whether there is a likelihood that a significant fraction of the (an) appropriate public will be confused. One of the few things that we know for certain about human perceptions is that even when faced with the same circumstances, the perceptions of different people often differ sharply. This difference in perceptions occurs even when the consumers involved all happen to be judges. That is, other judges’ perceptions of the hypothetical judicial robe trademarks might well differ from Judge Frank’s views. Moreover, experience shows that without actually asking those other judges, no one can realistically claim to know what their perceptions would be.
The essential issue of whether or not there is a likelihood of confusion in Lanham Act cases is: “How will a cross-section of people perceive competing trademarks and/or advertising?” The result is that a key piece of required evidence in most Lanham Act litigation is data on the perceptions of a cross-section of appropriate people. In other words, the variability in human perceptions dictates that it is not sufficient to call one or five or even ten witnesses; we generally need evidence from hundreds of people. It is for this reason that, in this author’s opinion, a ruling based solely on Judge Frank’s own reaction to the hypothetical trademarks he saw on two judicial robes would not, indeed could not, sufficiently show likelihood of confusion as required under the Lanham Act, but a competently conducted survey can.

Generally speaking, Lanham Act cases are concerned with what is likely to happen with regard to prospective purchases, rather than what has already happened. For instance, a typical false advertising claim states (or at least implies) that a cross-section of prospective consumers is or will be misled when they initially encounter the advertising in the real world. This need for simulating initial contact with the trademark or advertising demands that the witnesses must be “naïve” in the sense that they react as they would at first contact in real life. The basic value of a survey in Lanham Act litigation is that it scientifically approximates for the court the mental associations and reactions of prospective purchasers with the goal of clarifying for the court how the target population will perceive the specific trademarks and/or advertisements. In this sense, survey respondents are simply witnesses representing the target group of consumers. Their role is to provide evidence as to how that target group would respond.

There is no other way of developing this kind of evidence. That is, just as in voting research in which a survey is the only form of evidence that will clarify the perceptions of a cross-section of the voting public, so, too, in most Lanham Act cases, a survey simply is the only possible form of evidence that will clarify the perceptions of a cross-section of the buying public. Because there is

---

3. This forward looking attempt to predict harm before it happens is the underlying reason that likelihood of confusion, rather than actual confusion, is accepted in courts as evidence of infringement. Courts consider actual confusion as a factor in determining whether likelihood of confusion exits. This is both a forward looking attempt to predict harm before it happens as well as a means of keeping the standards for trademark infringement reasonable and attainable. Polaroid Corp. v. Polarad Electronics Corp., 287 F.2d 492, 495 (2d Cir. 1961).

4. Whether we call them witnesses, potential voters or potential consumers, all market research is about how some target group would respond in the marketplace.

5. Election day polling is only a very specific kind of survey, geared to tell you which candidate people prefer. Survey researchers use exit polls, still another form of survey, in order to understand why voters voted the way they did.
III. SURVEYS AS “IMPERFECT” EVIDENCE

To say that surveys are necessary does not mean that they are perfect instruments. Indeed, aside from their necessity, the most important thing to understand about Lanham Act surveys is that, just as is true for any survey, survey researchers understand that they are, at best, imperfect instruments for measuring human perceptions.7

In part, surveys are imperfect instruments because they operate under at least three types of unavoidable constraints:8

1. How people perceive advertising or trademarks in real life is never quite the same as how they perceive advertisements or trademarks in a survey environment. For instance, just the fact that respondents are aware that they are being surveyed is a clear and unavoidable compromise with “real marketplace conditions.”9

2. There are never infinite resources, so in trying to create the best possible survey one is always limited by the available time and money.

3. In many circumstances, respondents simply do not know the answer to the question we would like to ask.10 In such a case, we have to find some way of scientifically creating an inference as to how consumers would behave when faced with the problem in the real world.


8. This does not exhaust the constraints on real world surveys, but these particular problems arise in so many different ways that it is worth highlighting them.

9. However, to know you are being surveyed is not the same as knowing you are a witness, and, therefore, survey respondents remain “naive” in the sense discussed above.

10. A classic recent example was estimating the number of people who would desert Ralph Nader for Al Gore in the weekend before election. Ten days before the election, most people who preferred Nader did not know how they were going to behave on election day. So the Gore campaign had to utilize its resources to persuade such people that it was a “waste” of their vote to vote for Nader.
These unavoidable constraints lead inevitably to a succession of choices and compromises at every stage of the survey process. It is these choices and compromises that inevitably make the survey evidence imperfect. But even imperfect surveys, generally speaking, can and do clarify the issues in a case. In other words, surveys are clarifying evidence, not proof.

In this author’s opinion, it is this failure to distinguish between an inevitably imperfect compromise on the one hand, and a right or wrong decision on the other hand, that has occasionally led a court to hold a survey as being of “no value,” despite the fact that a reasonable person would recognize that the survey clearly shed some light on the issues.11

IV. SEEKING MODEL SURVEYS

Given the necessity for surveys in Lanham Act litigation, it is perhaps natural for commentators to seek to identify one or more model surveys as a standard way to provide a survey that is acceptable under the Federal Rules of Evidence. Responding to this challenge, commentators write articles that explicitly claim: “This is a (the) model survey.”12 Similarly, lawyers seek to rely on some standard survey, albeit in the less explicit guise of “Your honor, this survey is a ‘Squirt’ type survey and therefore should be credited.”13

In this author’s opinion, at its heart, the constant search for a model survey represents a refusal to accept the basic principle that any survey is the calling of witnesses in a specific case to testify about a specific set of “facts,”14 where those “facts” reflect the respondents’ (witnesses’) own perceptions. Once one accepts that presenting a survey in court is no more than the recounting of the

11. In their critiques of other people’s surveys, survey experts contribute to this view when they talk glibly of “fatal flaws.” A more nuanced critique would discuss how each claimed flaw in the design, implementation and/or analysis of a survey tends to distort rather than clarify how people would actually perceive the situation in the real world. See 5 McCarthy, supra n.1 at § 32:170-177 for descriptions of deficiencies in survey methodologies including location problems, irrelevant questions or exhibits, personnel and tabulation problems, bias problems, timing problems, copyright problems, insufficient number of respondents and leading questions. Some of these deficiencies in surveys stem from economic limitations which may be unavoidable (such as “[u]sing low paid, part-time, non-professional investigators” 5 McCarthy, id. at § 32:171), but McCarthy then states that “the majority rule is that while technical deficiencies can reduce a survey’s weight, they will not prevent the survey from being admitted into evidence.” 5 McCarthy, id at § 32:170.

12. One of the most recent examples of such an effort is William G. Barber, How To Do a Trademark Dilution Survey (or Perhaps How Not To Do One), 89 TMR 616 (1999).


14. As is the case for virtually this entire article, everything in this paragraph is of course true of any market research survey, albeit in most market research survey experts think of the respondents as potential customers rather than witnesses.
(hopefully clarifying) testimony of witnesses, why would anyone expect a model to exist? Surely, no reasonable litigator expects to use some model examination of an in-court witness, let alone a model cross-examination. Why then does anyone think there is, or should be, a detailed model examination of the respondents whom we have scheduled (of course, without their knowledge) to testify as witnesses? Another way to say this is that because a survey is no more than the calling of witnesses, exactly which witnesses to call (i.e., the universe and sample), and what to ask those witnesses (i.e., the questionnaire), always depends on the specific facts and legal theories of each case. Model surveys cannot hope to substitute for careful thinking about how underlying survey design principles apply to each new combination of facts and theory.

V. PRINCIPLES AND RULES

The case specific nature of all market research surveys makes it clear why, in the real world, the search for model surveys is bound to be fruitless. But, even if it is impossible to define a model survey, to maximize the weight accorded the survey evidence and to ensure that the evidence is admissible at trial, it might still be possible to identify standard “rules” to apply at various stages of the design, implementation or analysis of a survey.

However, as a survey researcher, this author has found that there are also few generally applicable rules. In part, as discussed above, this is because every survey design inevitably involves a series of compromises, and each of these compromises can by definition result in an “imperfection” in the survey. Therefore, to have the survey accepted by a court, each compromise needs to be defended in light of the specific facts and legal issues of that case. In other words, if surveys are to result in admissible and effective evidence, the need for making case specific design compromises implies that there is no absolute set of “rules” for carrying out the design, or implementing and analyzing a survey. Consequently, what is needed are principles to act as guides during the survey design process, rather than absolute rules that may be undercut by the unavoidable real world design compromises.

Throughout market research, the most important of the principles used in deciding the relevant universe to be surveyed and precisely who should be interviewed derives from the explicit recognition that compromise in the design and implementation of samples is inevitable. The issue is not whether the survey researcher managed to interview an ideal sample from an ideal universe, but rather whether the researcher managed to interview

---

15. Of course, there are principles that guide how a litigator develops an examination of a witness, and, correspondingly, there are principles that guide survey design and implementation. But principles are not the same as rules, let alone the same as “models.”
a “close enough approximation” to that sample. The basic sampling principle is that the survey researcher should seek to interview people who, if and when they were faced in a commercial situation with the trademarks and/or advertising at issue, could reasonably be expected to collectively have effectively the same reactions (i.e., within a reasonable approximation) as the collective reactions of actual purchasers and purchase influencers.16

This type of guiding principle for sampling, and similar statements for other aspects of survey design and implementation, makes it clear that, as is true for any survey, the issue in evaluating litigation surveys is to determine the likelihood that that particular survey will give a “close enough approximation” to the results that would be obtained from a perfect interview with a perfect sample. The underlying question is always to what degree does the survey actually conducted help to clarify (or obscure) for the court the specific issues of this case.17 In this light, the only way to actually decide if survey researchers have a reasonably close approximation is to provide the court with both the plaintiff’s and defendant’s answer to the following question:

Why do you think this sample (or questionnaire or analysis) would (or would not) give approximately the same answers to these questions as would be given by an ideal sample?

That is, just as the value of consumer market research is measured by the degree to which the results clarify the real world situation for the client, so, too, the value of a particular survey is not a question of whether it followed some detailed rules, but the degree to which the results are seen as likely to clarify the real world situation for the court. Principles, rather than rules, should guide the design of surveys and the weight to be accorded any survey cannot be decided on the basis of some set of absolute “rules,” but instead is a function of weighing the facts in each case. That is to say, survey evidence must be judged using the same kinds of arguments that characterize the assessment of the weight of any testimony from human witnesses.

VI. WHO TO INTERVIEW

A. Stages in the Sampling Decision Process

The first question that must be answered in any survey design is who to interview.18 A survey designer must consider the

---

16. See, e.g., 5 McCarthy, supra n.1 at § 32:159-160.
18. In pre-election surveys, for instance, the question of who to interview revolves around identifying “likely voters.”
following four stages in developing the sample of people interviewed:

1. The Universe(s)—For any given type of Lanham Act claim, the ideal universe is, in principle, specified by the law. Therefore to define the ideal universe in a specific case, the survey designer asks the question:

   “Who are all the people who constitute that segment of the population whose perception and state of mind are relevant to the issues in this case?”

For instance, ordinarily, the legally ideal universe for direct or forward (as opposed to reverse) trademark confusion is the potential customers of the defendant’s products and/or services, plus any individuals who influence the actual purchasers about the purchase decision. For reverse confusion cases, the appropriate survey universe is the senior user’s customer base. Thus, for a likelihood of confusion claim, the question defining the ideal universe becomes: “For this specific real product or service, who exactly are the potential purchasers and purchase influentials?”

2. The Sample Frame—Even though it is ordinarily possible to define the ideal universe, in practice this ideal is frequently no more than a theoretical concept of the desirable. To actually draw a sample, this ideal universe must be translated into a “sampling frame.” The sample frame means the answer to the question:

   “Regardless of how the universe is defined, who are all the people our sampling methodology allows us to actually talk to?”

It is important to recognize that, for a variety of reasons, the sampling frame is rarely better than a reasonable approximation of the ideal universe. It is consequently critical to specify principles that can be used to determine the degree to which the sampling frame is a “reasonable” approximation of the ideal universe when deciding “Who to Interview.”

3. The Sampling Procedure—A sampling frame provides us with a list (or some other mechanism) for specifying everyone who could be included in the sample. The actual sampling procedure then answers the question:

19. See 5 McCarthy, supra n.1 at § 32:159 and n.1.

"Of all those people in the sampling frame, how will we decide which ones we will actually try to interview?"

4. The Actual Sample—Which specific individuals are actually interviewed? It cannot be stressed too strongly that the quality of a sample is not determined by the people the sample designer specified to be interviewed. All that really counts is the specific individuals who consented to, and then actually completed, the interview.21

B. The Basic Principle of Sampling

One conventional legal definition22 of the appropriate universe in a trademark infringement trial is the defendant’s potential customers and any individuals who influence the actual purchasers about the purchase decision.23 Defining the potential customers is thus the first step in drawing a sample for a likelihood of confusion survey. However, for various reasons, the identity of these potential customers may not be clear. Indeed, it is not unusual for two survey “experts” to disagree as to who are the potential customers. While it is clear that such a disagreement should go only to the weight of the evidence and not its admissibility,24 some courts, unfortunately, have applied this principle in such a manner as to mean that a disagreement over the universe essentially renders the survey of little or no weight.25

A principle is needed to determine the degree to which the “improper” universe detracts from the reliability of the survey. It is never enough to simply say that a survey has “an improper universe.” It is always necessary to examine the degree to which one might expect that the differences between the universe of a

21. Unlike the first three stages, this fourth stage (actually agreeing to be interviewed) is primarily under the control of the respondent, although the designer certainly may influence cooperation rates in all kinds of ways such as paying for the respondent’s time.

22. See 5 McCarthy, supra n.1 at § 32:159-160, providing a definition of the appropriate universe and giving examples of appropriate universes in particular cases.

23. This article, unless otherwise specified, is limited to trademark confusion. While the principles discussed are similar for all forms of litigation surveys, the application of the principles can be quite different. For example, the definition of the legally appropriate universe for a trademark secondary meaning survey is generally different from that for a likelihood of confusion survey. See id.

24. 5 McCarthy, supra n.1 at § 32:161 (citing WGBH Educational Foundation, Inc. v. Penthouse International, Ltd., 453 F. Supp. 1347, 1351 (S.D.N.Y. 1978)). (“Disparity between the universe in the survey and ‘some optimal universe of people to whom such a survey should be directed reduces the evidentiary weight to be accorded its findings.’”).

25. For instance, in a case involving the names of camps, this researcher defined the proper universe as all parents of the appropriate age children, while the court felt that the universe should only have included parents of children with an established interest in soccer camps (the specific type of camp where confusion might occur). Seattle Storm, Inc. v. Ackerly Communications Group et al. (W.D. Wash. 2000) C00-74R.
survey and the “perfect” universe may alter the results and efficiency of the specific survey. 26

The principle for determining the weight to be given to the survey is the likelihood that the universe defined by the survey researcher will give statistically significantly different results from those that would be obtained from a survey of the universe the court considers relevant to the issues and facts of the case (i.e., the perfect universe). In a great number of cases, there is little, if any, reason to believe that using the universe as defined by the survey researcher, as opposed to the “perfect” universe, will result in any difference in results. If use of the “improper” universe will not change the results in any statistically significant way, then, from an evidentiary standpoint, the use of the “improper” universe should not detract from the value of the survey, and the weight given the survey should be essentially the same as the weight given a “perfect universe” survey.27

C. Surrogate Universes

However, even when the researchers and the court (or any other client) are in complete agreement as to exactly who the potential customers are, there are often still significant problems in specifying some way to tap into that audience, that is, in creating a sampling frame. In such cases, one, generally, is forced to define a substitute or “surrogate” universe.

By definition, using a surrogate universe is a potentially significant deviation from the perfect universe. But the key question is not, “Does this surrogate universe differ from the ideal?” Such deviation is the case by definition. Instead, the key question is “To what degree are the people in the surrogate universe likely to accurately reflect the views of the people in an ideal universe?” Following the principle introduced above, the legitimacy of a surrogate universe requires a situation-specific analysis as to whether the ideal and surrogate universes practically would produce the same results. If the results from the two universes are likely to be essentially the same, then the use of the surrogate universe is not fatal. Indeed, it does not detract from the value of the survey.

One commonly occurring situation, where the use of a surrogate universe is unavoidable, is when potential respondents

26. For instance, if one is attempting to market a new kind of convenience food, what is the effect of obtaining interviews with a disproportionately small number of the “busiest” people who may well be the best candidates for my product?

27. See, e.g., Iron Clad, L.P. v. Poly-America, Inc., 2000 U.S. Dist. LEXIS 10728 at 19-20 (N.D. Tex.), (Defendant objected to the use of a survey that did not eliminate non-members of Sam’s Price Club. Because neither side demonstrated that the trash bag market at Sam’s Price Club differed significantly from the general trash bag market, the court held that the survey had analyzed an appropriate universe).
literally do not know whether they are going to be potential consumers of the defendant’s product, so that neither the potential respondents nor anyone else can give an answer as to whether some particular individual is or is not likely to be a potential customer in some reasonable future time period. As an example, it is helpful to consider one possible definition of a universe to be used in a test of the likelihood of trademark confusion in the market for replacement refrigerators.\textsuperscript{28}

According to the conventional definition, the ideal universe for this case consists of those individuals who are likely to be in the market for a replacement refrigerator (e.g., a refrigerator purchased to replace one that breaks down) within some fixed time period (say the next year). But how many people know they will replace their refrigerator in the next year, or during the next two or three years? So what definition of a potential customer could we use to ascertain whether a specific individual is actually going to be a potential customer during the next year?

However, just because one cannot concretize the ideal universe, it does not follow that any surrogate universe will suffice. Instead, one must apply the appropriate principles to define and evaluate one or more surrogate universes. As discussed above, the appropriate guiding principle for defining a surrogate universe is people who will collectively behave in approximately the same way that actual purchasers are likely to behave under the same circumstances.

There are a number of universes that, to varying degrees, might be appropriate choices for a surrogate universe for replacement refrigerators. One can at least rank the appropriateness of various potential surrogate universes. For instance, three possible groups of people who would be in the market for replacement refrigerators are:

- People who have had to replace their refrigerator within the recent past (e.g., the last year). However, by definition, this surrogate universe immediately runs afoul of one of the general constraints on surveys, namely, that the incidence of such households for a reasonable period of time in the past is likely to be so low that we simply cannot afford to find enough eligible respondents.

\textsuperscript{28} While the example of a replacement refrigerator is chosen to make the concept clear, there are all kinds of purchases, ranging from everyday watches to Christmas gifts, where consumers have, at best, limited knowledge of their likely future decision to purchase the product or service. Moreover, it should be clear that this problem arises whether the reason for doing the survey is a concern over trademark infringement, or just a desire to sell refrigerators. This problem of choice of surrogate universe, as for virtually every other issue discussed in this article, is a standard one in survey research, not a situation unique to litigation surveys.
• Every person who owns a refrigerator. However, one might argue that there is a difference in the demographic makeup of all refrigerator owners versus owners of older refrigerators, which are more likely to be replaced, and this difference might affect how such persons perceive the trademarks.

• Those members of the general public who owned a refrigerator at least ten years old (or some other appropriate age). Assuming that the cost of finding enough people who own such refrigerators is not excessive, this may be the most appropriate surrogate universe.29

The replacement refrigerator example demonstrates the difficulty in specifying whether a particular individual is, or is not, likely to be in the ideal universe. A second set of problems in defining an ideal universe is that the pool of potential customers may change as a company expands its markets into new geographic areas or market segments.30 For instance, is the pool of potential customers limited to the geographic area in which a new trademark or product is being test marketed or does the pool include everybody in all those geographic areas where the product is currently sold, or perhaps all the areas where it might be sold in the future? Two contrasting examples demonstrate the range of situations where this can occur:

1. The defendant is introducing a new product, which has not been marketed before. Absent special circumstances, there is no reason to expect that the likelihood of confusion between two trademarks will be materially different in different geographic areas. Therefore, a very wide universe (e.g., the entire country) might be a perfectly acceptable surrogate universe.

2. The defendant sells its products only through Hispanic media and distribution channels. Here there is a convincing argument against defining too broad a universe. In particular, given the fact that monolingual English speakers might easily react differently to the trademarks (e.g., typical pronunciations might be very different), a proper surrogate universe would almost certainly have to be limited to Spanish-speaking or bilingual consumers.

29. But it is still a surrogate universe, which emphasizes that various surrogate universes are better or worse, not perfect or useless.

30. An analogous question is whether potential customers are limited to people with a previously expressed interest in a product or whether everyone qualifies who might reasonably come to want the product in the future.
D. Creating A Sampling Frame

The sampling frame consists of a list of all potential interviewees. In practice, there are three ways to develop such a list:

1. Develop a list of exactly those defined as eligible for the interview. For instance, for a study of the perceptions of doctors in a particular specialty, we might have available from state license records a list of all doctors in that specialty. Similarly, for a study among all the retail jewelers in a given geographic area, we might determine that the yellow pages provides a complete listing of such jewelers, and use the yellow pages as a sampling frame.

2. Develop a list that is smaller than the defined universe, and either
   - argue that for the purposes of the survey the differences are not significant; or
   - use some mechanism to supplement the list to more closely duplicate the ideal universe.

Telephone interviews and mall intercepts, the two usual ways of conducting a survey of the entire general public, exemplify two means of obtaining a sampling frame that is intrinsically smaller than the defined universe.

In the case of mall intercepts, clearly, we do not have an actual “list” of people who might be questioned at malls. However, since virtually all Americans go to malls, at least occasionally, the usual claim is that people who shop in malls make up such a large fraction of the general consuming public that the distinction is not meaningful, so that, for practical purposes, mall intercepts can be thought of as using a list of all Americans.

In the case of telephone interviews, no list of telephone households in the United States includes unlisted telephones. However, it is possible to begin with a list of listed telephone numbers and generate a list of all telephone numbers. Specifically, if one starts with all the listed numbers and adds the same small randomly selected digit to each listed number, the resulting list will include all telephone numbers (including many non-working or business numbers). For example, if the listed number is 234-5677 and the small randomly selected digit is three, the number dialed would be 234-5680. Because unlisted numbers are

---

31. As will become clear below, it is common to combine two or more of these methods in developing the sampling frame. However, it is clarifying to separate the steps involved.

32. There is another common objection to mall intercepts, namely, we do not know the relative probability of selecting different people for interviewing. This topic will be discussed at considerable length in the probability sampling section below.
distributed randomly among all telephone numbers, the result of this procedure is a list of all telephone numbers.33

3. Find a list that is larger than, but includes, the specified universe and then “screen” respondents to obtain data from only those who fall in the specified universe. Both mall intercept and random digit dialing interviewing are commonly used in combination with such screening procedures.

Using screening to establish the appropriate sampling frame frequently requires a compromise between a screen that perfectly matches the defined universe and the costs associated with completing interviews among such a low incidence group. Assume that the ideal (or best surrogate) universe involves a relatively small fraction of the general public. For example, assume that the out-of-pocket interviewing cost to find and interview 80 potential automobile lessees for 2002 would be $8,000.34 Assume also that because there are several times as many current or potential lessees of cars as there are potential lessees this year, finding and interviewing a total of two hundred current or potential lessees would also cost $8,000. Then for an expenditure of $8,000 for interviewing, we are, in effect, forced to choose between obtaining 80 interviews with the “ideal sample” and 200 interviews with a less than “ideal sample.”

From a statistician’s viewpoint this choice is best analyzed in terms of total error. The error from either sample is given by:

\[
\text{Total error} = \text{sampling error} + \text{sample bias}
\]

Sampling error occurs because we are interviewing only a sample rather than the entire population and it is essentially inversely proportional to the sample size (i.e., the sampling error declines as the sample size increases). Thus, sampling error occurs even for an ideal sample. Sample bias is the error that occurs because the sample deviates from the ideal sample.

In this case, the sampling error for the 80-person interview sample of potential lessees is larger than the sampling error for the 200-person interview sample of current and potential lessees, but the sample bias is smaller for the “ideal” potential lessee sample. The problem is that, while we can estimate the sampling error for both samples, we cannot get a numerical estimate of the sample bias for the sample of current and potential lessees (the sample bias for the “ideal” potential lessee sample is assumed to be zero).

33. Such so-called random digit dialing has other problems, notably, how to handle multiple telephones in the same household (a problem which has obviously been made much more intense with the advent of cellular phones). This issue will also be discussed in the section below on probability sampling.

34. These numbers are realistic estimates, based on actual costs of similar studies.
We can, however, make a reasonable argument about whether the perceptions about the trademarks at issue of people who are identified as current and potential lessees are likely to be significantly different from the perceptions of persons who are identified as potential lessees this year. Assume that for the trademarks at issue there is little, if any, reason to believe that the perceptions of the larger group (current and potential lessees) will be significantly different from the perceptions of the smaller group (potential lessees). In this author’s opinion, absent special circumstances concerning the trademarks, such an assumption often would be reasonable. However, under that assumption, the total sampling error of the 200 current and potential lessee sample will be less than the total error of the 80-person, potential lessee sample (i.e., the lower sampling error of the larger sample will more than compensate for its small or de minimus sample bias error). In such a case, not only is the 200-interview sample acceptable, but because it will give a lower total error, the less than “ideal” sample is actually preferable to the “ideal” sample, and should be used.

Generally speaking, screening can consist of pre-screening (screen questions asked before conducting the interview to screen in only those respondents who qualify for the interview), and/or post-screening (screen questions asked at the end of the interview, to allow separate analysis of those who qualify). This again involves a case-specific compromise. Pre-screening is somewhat less expensive than post-screening since it cuts the cost per screened-out individual (only the screen and not the whole interview is conducted). On the other hand, pre-screening raises issues of potential bias in that the screening process may alert the interviewee to the subject matter.35

E. Drawing the Sample—
The Myth of Probability Sampling

Once the sampling frame has been specified, the next step is to draw the actual sample of respondents. The ideal sample is conventionally discussed in terms of probability sampling.36 In surveys, a probability sample is defined as a sample in which each person in the universe has a known, non-zero probability of being included in the sample of actual respondents. In a well-defined strict projectability sense, probability samples are the gold standard in survey sampling. Unfortunately, there is only one

35. Post-screening frequently offers the additional benefit of permitting a comparison between two or more “interesting” groups of consumers (e.g., those who do and do not pass the screen). This comparison may be illuminating for the court.

small problem with probability samples defined in this way: *in the real world, almost no survey sample ever actually comes anywhere near to meeting the definition of a probability sample.*

If it true that the vast majority of real life survey samples do not qualify as probability samples, why then do researchers claim that many of the samples they use meet the standards of a probability sample? The answer appears to be that researchers tend to concentrate on what they can control, namely, the persons they (the researchers) select to interview. However, the full definition of a sample includes not just those whom the sample designer specifies to interview, but crucially, *it also includes the decision on the part of potential respondents as to whether or not they are willing to be interviewed.* Not everyone a survey researcher selects for an interview will be available and willing to be interviewed. As a result, the vast majority of real life samples grossly violate the definition of a probability sample in two critical ways:

- A serious proportion of those selected by the researcher to be interviewed will *not be available* to do the interview during the specified interviewing period. Such “respondents” actually have an *unknown* probability of being included in the results; and

- A serious proportion of those selected by the researcher to be interviewed will be contacted but *refuse* to do the interview. Such “respondents” actually have *zero* probability of being included in the results.

However, recall that the definition of a probability sample is that *all* members of the universe have a *known non-zero* chance of being interviewed, not just of being selected as potential respondents. Therefore, to be a probability sample, the great majority of people selected by the researcher to be interviewed must actually be interviewed. Taking into account people who are not available to take part in the interview and people who are available but refuse, current general public surveys claiming to be based on probability samples (e.g., random digit dialing telephone surveys as defined in the previous section), rarely attain actual completion rates even as high as fifty percent of the respondents specified by the researcher. Absent special circumstances, by any standard, a sample with a fifty percent or lower response rate does not qualify as a probability sample. As a result, unless one has the kind of resources available to the United States Census Bureau, few surveys ever actually come anywhere near to meeting the requirements for a probability sample.37

---

37. A few university survey groups also conduct surveys using what is, in effect, probability samples. However, this is the exception that proves the rule, because, generally
F. An Example of a Typical Sample

To clarify the issues involved in sampling, it is useful to consider an example from outside the world of litigation surveys: a survey of the dues paying members of a volunteer organization. That universe is reasonably well defined, and the sampling frame (i.e., a list of the dues paying members of the organization) is as close as possible to the definition of an ideal universe. Nevertheless, even in this situation, the sample cannot be called a probability sample unless we ignore the difference between the intent to interview and the actually completed interviews.

Assume some particular organization has 40,000 members and that we select every one hundredth person from an alphabetical list of the members to be interviewed. Then every member of the organization has a known, one percent chance (400 out of 40,000) of being selected for the survey. However, assume (as experience shows is clearly quite normal) the following:

- 20 of those 400 are away on business during the week that we allot for interviews;
- 8 of those 400 are in the hospital during the week that we allot for interviews;
- 72 of those 400 are too busy or too uninterested to spend the time to do a 20 minute telephone interview about this organization; and
- 100 of those 400 never agree to do interviews on any subject.

Under these assumptions, we will actually complete interviews with 200 of the 400 interviewees selected and unsuccessfully try to interview another 200 people. Of the 50 percent whom we did not interview, some had an unknown probability of being interviewed, depending upon whether they were available during the specific interviewing period (the not-at-homes), while others had no probability of completing an interview (the refusers).

Despite the fact that our 50 percent completion rate precludes considering those interviewed as constituting a probability sample, we have not yet fully addressed the usefulness of our sample. In speaking, only when the government finances the work does a university survey group have the necessary resources.

38. Even here the definition of the universe is subject to some debate. One of the many possible questions about the scope of the universe is should the universe include people who have been given a gift membership but actually have no real interest in the organization.

39. The example assumes that telephone interviews will be conducted. Sampling by mail, both in theory (because it allows for self-selection on the basis of the particular interview) and practice, is even worse. On the other hand, making appointments to see people and giving them a large honorarium on the spot (as is often done in surveys of physicians) will frequently result in much higher completion rates.
particular, the fact that we do not have a probability sample does not mean that we do not have a useful sample. Instead, what determines the usefulness of the survey is whether, with regard to the specific issues of the particular survey, there are likely to be significant differences between the 200 actually completed interviews and the 200 people who were supposed to be, but were not, interviewed. For the purposes of determining the usefulness of the survey, the only differences that matter (i.e., that can possibly be “significant”) are differences in the perceptions of the two groups with respect to each specific issue under study. Thus, we must examine the likelihood of such differences on a very detailed basis:

- If the subject of the survey is the travel services offered by the organization, then missing the 20 people away on business might (depending on how many other members travel frequently on business) have a significant impact on the results;
- If the subject of the survey is about the hospital insurance offered by the organization, then missing even just those eight people currently in the hospital might have a significant impact on the results;
- For almost any survey about attitudes about the organization itself, missing the 72 people who do not answer because they are uninterested in the organization will bias the results towards the extremes of favorability and unfavorability. Depending on the purposes of the survey, this may or may not matter;
- Most of the time, missing the next 100 people (those who refuse to do interviews) will not affect the results. However, suppose the survey is about alienation of our members from society. Is there a difference in the alienation from society of people who refuse to do interviews?

However, for the vast majority of issues, the actual 200 interviews will provide a realistic reflection of the views of the members of the organization; that is, they will be a useful sample.

G. Random Samples

This example of a sample drawn from an organization’s membership list makes it clear that if for no other reason than response rates in the real world, few, if any, samples can meet the strict requirements for a probability sample. But if probability samples are the gold standard, how can that be? Surveys are ubiquitous, and, as the vast majority of election polls show, often
clearly reasonably predictive. Is the predictive ability of surveys just a fortunate accident, or is there some underlying theory that says a probability sample is not a necessity for a useful survey?

The answer is that a probability sample is a sufficient, but not a necessary, condition for a useful survey. The major advantage of a probability sample is that it is the only way to be absolutely certain that a sample meets the necessary requirement for strict projectability. Strict projectability is, in turn, valuable because it allows us to define and calculate what is commonly referred to as the “margin of error” or, more precisely, as the “sampling error.” However, while strict projectability is a sufficient condition for a useful sample, generally speaking, it is not a necessary condition. The key word here is useful. Because we are not talking about perfection or certainty, but about realistic evidence, strict projectability, and, therefore, probability sampling is not necessary. In other words, for a sample to be useful, we do not have to be absolutely certain that it is perfectly projectable. We have to be only reasonably confident it is projectable to a reasonable approximation.

One way to increase the probability that a survey sample will be projectable is to draw the sample using a random procedure to select respondents. By definition, a random selection procedure is one in which the method for selecting respondents (including the self-selection involved in agreeing to be interviewed) is not statistically correlated to any meaningful degree with the issues being examined or tested by the survey. If a selection procedure is totally uncorrelated with the issues being examined, then the sample is projectable to all those people who had a non-zero chance of being selected into the sample (e.g., if mall intercepts are used, the sample is projectable to people who shop in malls or more accurately to people who shop in malls and do not flatly refuse to do interviews). If the selection procedure is not statistically correlated to any meaningful degree with the issues being examined, then, by definition, the sample is a good approximation to a perfectly projectable sample.

Fortunately, in practice, a great many survey samples (and, in particular, trademark litigation survey samples) seem overwhelmingly likely to meet this requirement of no meaningful correlation and, therefore, come close enough to projectability to make the survey useful. For instance, for most organizations, using the list sample approach discussed above, we would expect

---

40. In any case, the margin of error is a much overused indicator of the quality of a survey. People tend to think of the margin of error as reflecting all the errors in a survey, when, in fact, it refers only to the error due to the fact we have a sample rather than a census. In surveys of human populations, other errors are often potentially much larger (and thus more important) than the sampling error (e.g., the error as a result of refusals is one particularly pertinent example).
few, if any, significant differences in perceptions between the 200 completed interviews and the 200 people who were selected but not interviewed on many issues. For such issues, the sample of people actually interviewed becomes a close, and therefore usable, approximation of a projectable sample.

Because it is so often desirable to show products to respondents, mall intercepts are the most important class of usable random, non-probability, sampling for litigation surveys. The key question then becomes: “Is it likely that the process by which people are selected in a mall interview is likely to be statistically correlated to any meaningful degree with the issues being examined?” Of course, this question needs to be thought through in each specific case. However, for many litigation issues, such as the likelihood of confusion between two household product trademarks, it is hard to see any reason why such a correlation would exist.

For example, suppose we are interested in measuring the likelihood of trademark confusion between two brands of baked goods. In such a case, it is hard to see why randomly stopping and interviewing adults responsible for their household food shopping in a mall would yield results that would differ in any meaningful way from a perfect sample of all adults responsible for baked goods shopping. Therefore, it is reasonable to claim that such mall intercepts are projectable, at least to a level clearly usable as evidence in a civil trial.

A “convenience” sample of small businesses is another example of an important class of trademark litigation surveys where the sample is typically based on methods that are, strictly speaking, non-projectable, but are very likely to meet the real requirement of no meaningful correlation. Such a convenience sample is typically defined as businesses in a few office and/or industrial parks close to the office of the interviewing firm. Where owners or managers of small businesses are being interviewed about some general business service (e.g., telephones), it is hard to see why such a convenience sample will ever give results that differ in any significant way from the results that would be obtained by a perfect random sample of small businesses selected from across an entire metropolitan area.

Moreover, in most cases, interviewing across an entire metropolitan population is considerably more expensive than interviewing in a more limited geographic area. Therefore, for the same total cost, the convenience sample will produce more interviews and, therefore, a lower “sampling error.” Because there is no reason to expect any bias to be introduced by the use of the “conveniently located businesses,” while the larger possible sample (at fixed cost) will produce a lower sampling error, this is actually an example of where the less than ideal sample will generally yield
a lower total error than the “ideal” sample. That is, the “less than ideal” sample (i.e., the convenience sample) is actually preferable to the “ideal” sample (i.e., the random metropolitan area-wide sample) and the convenience sample should be used.

VII. WHAT WAS THE RESPONDENT'S TASK?
THE CONTENT OF THE SURVEY

A. Principles Underlying Content Design

Even though there are problems in implementing a real world sample, there often is at least a theoretical way to obtain an ideal sample. Nonetheless, as we have seen, because actual sample design raises case specific problems, virtually every actual sample involves a series of compromises. Each such compromise is by definition an “imperfection” in the survey.

If such imperfections are ubiquitous, even in the design of samples, it is surely not surprising that they are inevitable in the design of the content of the survey (where no theoretical ideal exists).

Indeed, in designing the content of any type of survey, there are few rules. Instead, the design process is primarily the case specific application of a few general principles. There are three basic principles that affect the design of a survey’s content:

- The “scientific” principle, i.e., that, ideally, the researcher will have complete control over what is shown to and/or asked of the respondent;

- The “reality” principle, i.e., that the researcher wants to come as close as possible to marketplace conditions. One critical way that surveys deviate from reality is that respondents know they are taking part in a survey rather than an actual purchase situation. Finding ways to minimize such “survey artifact” effects is often the key to effective content design; and

- The “error measurement” principle, i.e., the need to build in to any design a way to estimate the magnitude of the various sources of error, because in any survey design there are always at least potential sources of error and, in particular, errors arising from the gap between the survey process and the real marketplace (survey artifact errors or noise).
B. The Phases of Content Design

In discussing how these principles can be applied to the content design of a specific survey, four different design phases can be distinguished, although the separations between these phases are hardly precise. In particular, the researcher often has to go through several iterations of the last two phases:

1. Selection of an overall approach;
2. Selection of the interview structure;
3. Specification of the questions including the question sequencing; and
4. Selection of “control” stimuli.

C. The Overall Approach

The first step in the content design process is the selection of the overall approach. The most commonly used overall approaches are by telephone, in person at a place of business, mall intercepts, and shopper studies.41

1. Telephone—If respondents do not need to see materials,42 telephone interviewing is commonly the most cost effective method.

2. Business in person—If respondents need to see materials, and the issues involve purchase decisions made routinely by managers or professionals in the course of conducting their business, sending the interviewer to the individual’s office or place of business is generally the most cost-effective approach.

3. Mall intercept—If the appropriate universe is all (or at least a major subset) of the general public and respondents need to see materials, then mall intercept interviewing (perhaps including screening to reach the desired subset) is generally the most cost-effective approach.

41. Although conditions are rapidly changing, at least currently Internet interviewing raises serious problems in self-selection, and thus will not be considered further in this article. A number of other approaches have also been occasionally used. Two of these—mail interviews and trade show in person interviews—suffer from serious sampling problems and consequently are rarely used in litigation. A somewhat more usable method is a combined telephone/mail approach in which respondents are recruited by telephone, then mailed materials and re-interviewed by telephone. The difficulty here is the lack of any way of insuring that the respondent has not opened the mailing in advance. Cf. Quality Inns Int’l v. McDonald’s Corp., 695 F. Supp. 188 (D. Md. 1988) (an example of a survey in which mail-out with telephone follow-up was rejected by the court).

42. However, in Lanham Act cases, the need to show material to respondents is so common that telephone interviews of the general public play a much smaller role in litigation surveys than they do in other common uses of surveys.
4. Shopper—In a shopper survey, interviewers enter a store or similar establishment and act as if they are “shopping” for a particular item. If someone in the retail store where an item is normally bought (i.e., in the real marketplace) takes an active part in the decision making process\textsuperscript{43} a shopper survey can be an effective approach.

However, emphasizing once again that surveys provide evidence, not proof, even in the initial choice of an overall approach, there is an unavoidable and significant compromise. This compromise is required because, generally speaking, two of the principles introduced above are almost always in conflict:

- On the one hand, the questioner of witnesses (i.e., the researcher) would like to be able to determine everything with which the witness (i.e., the respondent) is presented. In particular, the researcher would like to completely determine\textsuperscript{44} the stimuli and the exact wording and ordering of all questions asked.

- On the other hand, because the Lanham Act is concerned with avoiding confusion in the real marketplace, it is obviously important that the interview situation come as close as possible to replicating the real world.\textsuperscript{45}

In the great majority of situations, a shopper survey design is the only available way for researchers to make this compromise in favor of marketplace reality. That is, almost always, a shopper survey is the only way to gather information where the “respondent” is in a common everyday situation such that the “respondent” does not know they are actually participants in a survey. Unfortunately, the mechanism by which shopper surveys provide a realistic replication of marketplace reality involves a real cost. This cost arises from the fact that in a shopper survey the researcher is usually able to control only the first question asked. Once that first question has been asked, usually the best the researcher can do is to provide the interviewer with a guide as to how to behave. Thus, in exchange for a shopper survey’s mimicry of real marketplace conditions, researchers must give up a

\textsuperscript{43} The requirement that a retail clerk take an active part in the decision process means that the necessary conditions for a shopper survey occur in only a small fraction of Lanham Act cases. See Lon Tai Shing Co. v. Koch + Lowy, 19 U.S.P.Q.2d 1081 (S.D.N.Y. 1990); Thomas & Betts Corp. v. Panduit Corp., 138 F.3d 277 (7th Cir. 1998).

\textsuperscript{44} The reader may note some awkwardness in this phraseology. That awkwardness can be reduced by using the phraseology, “the researcher would like to control the process.” Unfortunately, experience shows this can lead to confusion between the researcher controlling the process and the researcher using “controls,” where the term “controls” has a specific technical meaning to be discussed at some length below.

\textsuperscript{45} See 5 McCarthy, supra n.1 at § 32:159.
significant proportion of their ability to control exactly what is to take place during the interview.

There are also significant problems associated with all of the general approaches that permit the researcher to define the questionnaire and interview procedures. Such problems are inevitable in any situation in which respondents know they are respondents. By definition, simply being aware that they are taking part in a survey removes respondents from the real marketplace and places them in a somewhat artificial situation. Given that all respondents are in an artificial situation, it is always possible that some will react differently than they would in the marketplace. Differences between how people react in the real marketplace and how they react when they know they are in a survey are deviations from reality that are caused by the survey. These deviations, or survey artifact effects, are commonly referred to as “noise.”

D. The Interview Structure

The interview structure is the manner in which stimuli are shown to the respondent. Experience shows that different interview structures can be more or less likely to lead to guessing. As a result, and as is true for almost every aspect of interview design, the choice of structure has a direct effect on the level of guessing or noise. This, of course, is completely consistent with the idea that a primary cause of noise is that each respondent is making an attempt to “pass the test.” Thus, if the structure used is designed to provide respondents with alternative ways (i.e., other than with respect to the issue of interest) to feel they have “passed the test,” the researcher will reduce the tendency to guess and thus the level of noise.

There are a wide variety of fairly general options available to the designer in the interview structure. The following are a few examples of interview structures to be used in the most straightforward likelihood of confusion situations.46

1. Either: (a) all stimuli are shown to respondents in the same room (a one-room methodology), or (b) only the plaintiff’s mark is shown in the first room, with the remaining stimuli shown in a second room (a two-room methodology).47 Using a one-room methodology, each of the stimuli are equally “important,” while using a two-room

46. This wide variety of alternatives in interview structures means that selection of the general structure is the point at which any effort to define a model survey begins to break down.

47. A similar alternative is to show only the defendant’s mark in the first room, in which case the plaintiff’s mark and all of the controls are being compared to the defendant’s mark.
methodology, all the marks shown in the second room are essentially being compared to the first room mark. If it is the plaintiff’s mark that is shown in the first room, then in the second room the defendant’s mark and all the controls are being compared to the plaintiff’s mark.

2. Whether in a one-room test or in the second room of a two-room test, all of the stimuli can be shown together or the stimuli can be shown sequentially.

3. If all of the stimuli are shown together, products in the class of relevant goods can be presented in a number of ways. For instance, they may be set out along side each other, (i.e., in an array), or they may be displayed in a manner that replicates the presentation of the goods on shelves in a store with a variety of products. Alternatively, respondents can be asked to perform some form of sorting exercise, using file cards displaying names or photographs of products.

4. Depending in part upon which of the above interview structures is used, there are various ways in which the trademarks can be presented to a respondent (e.g., orally or visually, and if visually, by presenting products or pictures of products).

E. Deciding on an Interview Structure

Deciding upon the specific interview structure to be used involves balancing competing interests of costs and validity. The “best” balance is driven by the facts and legal theories of each specific case. Because essentially every case raises different issues, an exhaustive discussion is beyond the scope of this paper. However, it is helpful to examine a few of the tradeoffs that may arise when deciding between a one-room or two-room methodology.

How to construct an interview structure in any given case involves considering the question: “Under what circumstances is it claimed that there is a likelihood of confusion?” If the claim is that confusion will arise when the products are seen side by side, then the one-room test clearly comes closer than the two-room test to replicating actual market conditions. Ordinarily, however, confusion between product names is more likely to arise when the plaintiff’s product is first seen outside the retail environment, (e.g., after the sale has been made at a friend’s house), and then the defendant’s product is encountered by the consumer in the marketplace. If such a sequential situation is likely to reflect reality, then presenting the plaintiff’s product in the first room of a two-room methodology and presenting the defendant’s product along with the control products in a second room clearly approximates the real world better than a one-room test.
On the other hand, using a one-room test it is often simpler to design an interview structure that reduces the noise by providing an “easy task” for the respondent. For instance, in the one-room test, the control stimuli can easily be compared to each other (as well as to the plaintiff’s and defendant’s marks) and thus it is feasible to have an “easy task” consisting of a “clear” connection between two of the control stimuli. In contrast, a two-room test involves a comparison of all of the marks to the plaintiff’s mark and, as such, there often is no easy, equivalent design procedure available in a two-room methodology.

F. The Questions—

The Respondent as Witness

As emphasized throughout this article, each respondent to a litigation survey is a witness, so that the questions posed to each respondent constitute the examination of a potential consumer as a witness. Development of a good witness examination is not an area of expertise limited to survey researchers. Indeed, the survey researcher who questions respondents in the field, and the litigator who examines witnesses in court, are engaged in much the same process. The only real difference is that witnesses in court know they are under oath and that as witnesses they are subject to cross-examination. Thus, in court, witnesses are subject to substantial pressures to tell the truth In contrast, survey witnesses may feel pressure to provide an answer to every question (i.e., to guess), and therefore, noise is primarily a problem with survey respondents.

All of the other possible sources of error in questioning any human being are similar, whether the witness is being examined in court or interviewed in a survey. For example, open-ended questions, where the answer categories are not defined by the examiner, raise similar problems, whether asked in court or in a survey context. In particular, the answers given in response to open-ended questions may not be (precisely) responsive to the issue of interest. An ambiguous answer is an ambiguous answer, whether it comes from a survey respondent or a fact witness on the stand. On the other hand, in both situations, it is true that the answer to an open-ended question will often make clear how the examinee understood the question.

Conversely, a question in which the witness is asked to choose from pre-specified answer categories can generally be designed to force the witness to focus more precisely on the issue of interest. However, closed-ended questions raise a variety of other possible types of error. As one example, unless all the reasonable answer
categories are included, closed-ended questions can give seriously misleading results.48

Errors arising from the pressure to guess are far from the only sources of noise that may arise from the design of a survey interview structure. However, aside from the risk of a tendency to guess, examining in-court witnesses and interviewing respondents in survey circumstances raise much the same issues in question and questionnaire design. As there is ample literature on direct examination and cross-examination of trial witnesses, this article will focus on the specific issues involved in minimizing the noise caused by guessing. That is, the article will distinguish noise from other kinds of witness error and define noise as limited to those errors that result from pressure arising from the respondent's knowledge that they are taking part in a survey.49

G. The Sources of Noise

Before discussing ways in which to minimize noise, it is important to consider two ways in which noise can be caused whenever respondents know they are being surveyed. The first is a type of “leading” question that suggests an answer (presumably the answer “desired” by the researcher).50 However, it also is leading to suggest to a survey-respondent that there is a correct answer to a question.

The great majority of people take for granted that they would not be interviewed unless they were expected to be able to answer at least some of the survey questions. From the respondent’s viewpoint, why else would the researcher conduct the survey? In other words, just because they are being interviewed, most respondents assume that at least some of the questions in a survey have “an” answer, and moreover, that they should know the answer. Accordingly, many people treat surveys as a “test,” and accordingly guess in an effort to show they are not stupid and do “know” the answer.51

48. A classic example was the United States Census Bureau's insistence (prior to the 2000 census) that every person had to specify exactly one single “race” (i.e., there was no way on the census form to claim to be both white and black, or to be an Asian-African American).

49. For example, in a shopper survey, because the respondent (the retail clerk) is unaware that a survey is being conducted, by definition, there is no possibility of noise. The answers always reflect what would happen in the real marketplace. Nonetheless, there are other potential causes of error (e.g., the desire of the clerk to make a sale). Thus, even in shopper surveys, controls are still necessary to clarify which answers reflect legally meaningful indicia of consumer confusion.

50. 5 McCarthy, supra n.1 at § 32:172.

51. The classic demonstration of this was a study by researchers from the University of Cincinnati. About seven percent of the respondents in their survey said they had written their congressman about a specific bill, which was interesting considering the researchers
A second closely related source of noise is that researchers frequently ask questions that might never be considered by the respondent in the real marketplace. To the degree that respondents answer these questions anyway, then in the specific sense that such perceptions would not engender any actual behavior, the perceptions measured would not exist in the marketplace and are not based on any actual behavior.

**H. Minimizing Noise—Telling Respondents Not to Guess**

In trying to minimize survey artifact effects or noise, researchers have tried to develop global design techniques that may be used regardless of the structure or general approach of the interview. Before turning to some useful global design survey techniques, this article will discuss one technique that the author thinks is vastly over-rated.

It is frequently claimed that giving a respondent an initial instruction “not to guess,” or “not to be afraid to say they don’t know” will, in fact, reduce the tendency to guess. Irrespective of the apparent logic, in the real world, such an admonition generally does not have any discernible effect on respondents.\(^{52}\) This absence of an effect is an outgrowth of the underlying reason for a respondent’s guessing, namely, that respondents assume that there are answers to the questions posed. Given this assumption, an admonition “not to guess” is as likely to confirm for some respondents that there is a “right” answer that they should be smart enough to know, and thus to encourage guessing, as it is to discourage guessing among other respondents.

This general admonition against guessing sometimes is manifested as an insistence upon the value of including as an answer category to individual questions: “I don’t know.” Including an “I don’t know” option is, in this author’s opinion, less likely to encourage guessing and thus may be a harmless practice, although not of real value.

---

\(^{52}\) An example is the fairly well-known “Brooklyn Dodgers” case. Major League Baseball Properties v. Sed Non Olet Donarius, Ltd., 817 F. Supp. 1103 (S.D.N.Y. 1993). Here, based on my criticism, the plaintiff repeated their survey introducing a “Don’t know” option at critical points. The results of the new survey were the same as the results of the original survey. A related phenomenon is the ability of human beings to ignore even explicit disclaimers. J. Jacoby and G. Szybillo, Why Disclaimers Fail, 84 TMR 224 (1984).
I. Minimizing Noise—
Some Useful Questionnaire Methods

One technique that does seem to reduce noise is for the researcher to describe to the respondent the substance of some, or all, of the questionnaire. For instance, telling the respondent that the same question is going to be posed about a number of different stimuli tends to reduce the pressure to get an answer to any one stimulus “correct.” This reduces the pressure to guess. In the case of survey stimuli involving cards or pictures, this message can be strengthened by shuffling the cards or pictures in front of the respondent, thus reinforcing the respondent’s understanding that no specific stimulus is particularly important.

A second approach is to pose a very “easy” question at the very beginning of the interview prior to asking questions about the stimuli at issue in the case. If a respondent knows that they have gotten one question “right,” this has the effect of reducing the pressure to guess later in the interview. A related way to reduce noise is to divert attention from the stimulus of interest (e.g., the defendant’s trademark), for example, through the judicious selection of some additional stimuli. For instance, an “array” type of structure (i.e., a display of the plaintiff’s and the defendant’s products plus four or five control products side by side) may be used to measure the likelihood of confusion between the product configuration trade dress of physically small products.

These two approaches may be combined by making the “easy question” concern some “clear” connection between two of the “control” stimuli. For example, in a color/shape trade dress case, the researcher may use two “control” stimuli that are quite close in shape and color to each other) thus reducing the pressure on respondents to guess about any other connections between the plaintiff’s and defendant’s marks.

J. Measuring Residual Noise—
“Control” Stimuli

The need to minimize noise is an important factor that impacts every aspect of a survey design. Nevertheless, noise is always at least potentially present in even the most sophisticated and complete design. Therefore, to estimate what would occur in the real marketplace, we must provide an answer to the question: “How can the court distinguish between the perceptions of consumers concerning these trademarks or advertisements in the real marketplace and those perceptions which are generated by the survey process itself?” As a result, we must always provide a way

---

53. This is analogous in form, but different in purpose, to the use (to be discussed below) of additional stimuli as measures of residual noise (i.e., as controls).
to measure the residual noise.\textsuperscript{54} If we can find a way to measure or filter out the residual noise, then we can simply subtract the measured value of noise from the measured value of the issue of interest to get the net of noise best estimate of the issue of interest. For instance:

\[
\text{Best estimate of likelihood of confusion} = \text{Measured likelihood of confusion} - \text{measured noise}
\]

Researchers measure residual noise through the use of controls.\textsuperscript{55} To demonstrate how controls can be used in litigation surveys, it is useful to consider two analogous situations: a) determining through the use of a line-up whether a given suspect is likely to be the actual perpetrator of a crime, and b) using a placebo (e.g., a sugar pill) to test whether a new drug or medical procedure is an actual improvement over existing practice. Controls in surveys are analogous to introducing people who are not suspects into a criminal lineup or to giving a “placebo” to some fraction of the patients in testing of a new drug. Indeed, the use of multiple people in a lineup and placebos in drug testing are essentially the same response to the same problem as are the use of controls in consumer surveys. That is, many people, when they know they are in some form of test or experiment, feel pressure to provide the right answer. That is true whether the perceived test is “Is this the person you saw that night?”, “Did that drug cure your illness?”,\textsuperscript{56} or “Are these two trademarks from the same source?” Just as, in most circumstances, no one would trust the results of a “lineup” in which witnesses were shown only the suspect, the results of a litigation survey, without any controls, are strongly suspect.

\textbf{K. Selecting the Controls}

The analogy to criminal line-ups and medical experiments can be used to suggest some more explicit control guidelines.

Suppose that the one thing all of the witnesses to a robbery agree on is that the perpetrator was a middle-aged white male (perhaps he was even caught on a tape, but one that did not show his face clearly). Then showing witnesses the suspect, a Chinese-American girl, a black man, and an elderly white woman is not

\textsuperscript{54} Of course, occasionally the measured residual noise turns out to be nil.

\textsuperscript{55} While well-designed controls may help to minimize the noise, the primary purpose of controls is that they are the way we measure the residual noise. See, e.g., Volkswagen Astiengesellschaft v. Uptown Motors, 1995 WL 605605 at 3 (S.D.N.Y.) (“The need for controls is sometimes addressed as a need to screen responses for background noise, in other words, to eliminate those respondents who are confused by something other than the presence of the trademark at issue.”).

\textsuperscript{56} The fact that generally a higher proportion of people will get well (i.e., lose their symptoms) even when given the placebo, than when they are not in a drug trial emphasizes the pressure of this suggestion.
really a proper line-up. On the other hand, as long as most of the members of the line-up are middle-aged white males, their faces may be quite different. To go one step further, if all the witnesses agree that the perpetrator had red hair, then only redheads should be included in the lineup. There is no reason that one or more of the people involved should not be wearing red wigs. Indeed, it is a certainty that any witness picking out the cop with the red wig is guessing.

Similarly, a sugar pill may be a useful control in a test of a new analgesic pill, but such a placebo is of little use in a test of using a new barium formulation for taking an MRI of the lower intestine. However, generally speaking, the current barium formulation will be an excellent control for the new formulation, even though it is no longer, technically speaking, a placebo.

The general analogy to line-ups and placebos in medical tests, together with these specific examples, suggests four principles for choosing the controls in any Lanham Act survey:

1. Controls do not have to exist in the marketplace but can be formulated for the selected survey. This is because we are not trying to measure whether the controls are “infringing,” but whether people are guessing in the survey;

2. Multiple controls are generally, but not always, better than a single control;

3. At least some of the controls have to provide reasonable “guesses” in answer to the particular question of interest. That is, if in a particular test, a control stimulus is so different that no respondent would ever guess it, even when it is the only stimulus shown, then such a control cannot measure whether people are guessing with response to the stimulus of actual interest; and

4. The best control may be the defendant’s previous product or advertisement. One has a fairly strong argument that there is a strong likelihood of confusion if significantly more people are confused as to the source of the defendant’s new packaging than were confused as to the source of the defendant’s old packaging. However, this principle may come into conflict with principle three, and in most such situations, the requirement of reasonableness must still be satisfied. This can often be achieved by using multiple controls, with the defendant’s previous product or advertisement as one of those controls.
VIII. CONDUCTING THE SURVEY

Having selected the sample and designed the content of the survey, what remains is implementation. This article will consider four implementation issues.

A. Double Blinding

Double blinding means that neither the respondent nor the interviewer should know the purpose of the survey. The need for double blinding arises from the impossibility of anyone being able to guarantee that interviewers will be able to control their body language or other non-verbal cues. An interviewer who knows the purpose of the study may inadvertently (or perhaps in an intentional desire to please the client or boss) indicate to the respondent what is the “desired” answer. It is equivalent to the double blind requirement in a drug or medical device study that neither the patient nor the treating physician know which patients are getting the test drug and which are getting the placebo.

Because there is no possible way to measure its effects, a failure to double blind is one of the few violations of an ideal survey design that truly can be considered a fatal flaw. If the survey is conducted “in house” by interviewers directly employed by the researcher’s firm, rather than through a completely insulated outside interviewing firm, it becomes much harder to guarantee double blinding. In this author’s opinion, for that reason, any survey done in-house is presumptively fatally flawed.

B. “Knowledgeable” Respondents

Many survey researchers screen out what they define as “knowledgeable” respondents by asking questions such as: “Do you or does anyone in your household work for an advertising agency, a marketing research firm or a manufacturer, distributor, dealer or retailer of (whatever is the business of the plaintiff and/or defendant)?” In this author’s opinion, such screening is not only costly and inefficient, but if it actually does eliminate “knowledgeable” respondents, it introduces a serious and undefendable bias into the sampling process.

The screening out of knowledgeable respondents is inappropriate both for practical and legal reasons. The practical reason is that it is absurd to claim to be able to determine who is likely to be “knowledgeable” about trademark issues in a specific case (which is the only knowledge of interest) through this type of screen question. To show its absurdity, consider that in almost all consumer product cases asking if anyone in the household works for a manufacturer, distributor, dealer or retailer in whatever is the business of the plaintiff and defendant may screen out the wife of a night watchman for a Wal-Mart as potential respondent, while
leaving as potential respondents lawyers who make their living litigating trademark actions. Given the impracticality of effectively and economically identifying “knowledgeable” respondents, it is fortunate that there seems to be no legal basis whatsoever for eliminating the truly more knowledgeable. This author is unaware of any legal argument, either in the Lanham Act or the related law, that speaks to a likelihood of confusion among a more or less knowledgeable public. The criterion for trademark infringement is likelihood of confusion among all potential consumers. This criterion, if anything, indicates added interest in the more knowledgeable consumer, because people who buy a given product or service will presumably be more knowledgeable about what they buy than will the non-consuming public. Indeed, in the logically extreme but common situation where the product is sold only to professionals and/or to people in a specific business, screening out those who work in the industry of interest would screen out everyone who ever purchases the product, and there would literally be no one left to interview.

C. Interviewer Instructions

Researchers differ greatly on what is considered the appropriate level of interviewer instructions detail. In reality, all of the litigation survey interviewing should be conducted by trained, experienced interviewers. In this author’s experience, litigation survey questionnaires are almost always the shortest and least complex questionnaires done in any area of survey research. Of course, this does not mean that, on occasion, interviewers either may misunderstand the task or even intentionally cheat. What it does say is that such misunderstandings and/or cheating rarely, if ever, can be affected by lengthier instructions. The truth is that, while hiring experienced interviewers from reliable interviewing firms helps to avoid the risk of misunderstanding or cheating, it is impossible to accomplish risk-free interviewing.

D. Validation

After the interview is completed, validation should be done to check whether the interviewing was properly carried out. In this author’s opinion, one of the myths of survey research is that the purpose of validation is to check on interviews. It is not. Rather, the purpose of validation is to check on the interviewers. This is to ensure that they actually did the work and that they did it correctly.

There are two main methods for validating interviewers. The conventional method is to ask respondents at the end of the
interview for their first name and telephone number, and then to have an independent interviewing firm call a sample of respondents to confirm that they had in fact been interviewed. This validation call-back can only provide a check on whether or not the interview was actually conducted at all because it ordinarily includes, at most, one question about the general subject matter of the interview. In particular, if the interview was done incompletely or if any, or all, of the interview was conducted improperly, this general method of validation is of little use.

In a like manner, because we are trying to validate the interviewers and not the interviews, it misses the point to set a very high standard, such as attempting to validate 50 percent or even all of the survey respondents. Depending on how many interviews each interviewer conducted, it is possible to ascertain whether an interviewer actually conducted the interviews with a random sample of perhaps 15 to 20 percent of each interviewer’s work. Similarly, it does not show that the interview was not actually done if a respondent gave a false telephone number, or denies having done the interview when called back for purposes of validation. Such a response may simply mean that the respondent, for whatever reason, does not wish to be bothered. All that is required is that an additional portion of that interviewer’s work be validated to demonstrate that the interviews were actually conducted. Indeed, if a researcher really believes that an interviewer signed for even one interview that was not actually conducted, the researcher has no choice but to discard all of that interviewer’s work. Otherwise, the researcher would be accepting part of the work of an interviewer believed to be a cheat. It seems impossible to testify under oath that such work is to be trusted.

Because call-back validation can only validate whether interviews were conducted at all, this author believes it is important to carry out a second form of validation. This second form of validation consists of arranging all the interviews of a particular interviewer in chronological order and having an experienced analyst review them. Such procedures can often identify, not just whether the interviews as a whole were done, but whether the interviewer followed instructions, and conducted the entire interview. Moreover, following this procedure adds such depth to the analysis of the results that it is surely worth doing on that basis alone, with the validation value being an added bonus.

IX. CONCLUSION

Because it is necessary to introduce as evidence the perceptions of a cross-section of an appropriate public, surveys are a necessity in most Lanham Act litigation. However, because they are inherently imperfect measures of human perceptions, surveys
do not in themselves constitute proof, but rather constitute evidence introduced to support specific legal theories of the case. Furthermore, because the development of a litigation survey is so case specific, there are no model surveys and few rules. However, there are a small number of relatively general principles that can be used to guide the creation, development and completion of a survey. The underlying theme is that survey respondents are essentially a special class of witnesses who are brought before the court through the survey and the expert’s testimony in order to show how an appropriate cross-section of consumers of the particular product or service will perceive the stimuli at issue. No one survey performs this task perfectly. But imperfections in surveys do not render survey evidence useless. The value of a survey to the trier of fact is rarely simply a yes or no proposition. Thus, the major question in analyzing the value of a survey remains: “Does this particular group of witnesses being shown these particular stimuli, and being asked these particular questions, provide evidence that will clarify the issues for the court?”