INTRODUCTION

In the social and behavioral sciences, the term pretest often refers to a preliminary small-scale investigation of the proposed questionnaire, while the term pilot test includes more: a small-scale investigation that tries out survey parameters in addition to the questionnaire design, including the screening procedure, the determination of likely incidence and response rate, the choice of survey modality, and so on. However, the terms are often used interchangeably to describe preliminary investigations of any kind aimed at trying out any aspect of a survey that is being considered as extrinsic evidence in trademark or deceptive advertising litigation. In this chapter, the term pilot (test/work) will be used in this broader sense as including preliminary empirical investigations of any kind.

1. A pretest is defined as a “trial run of a questionnaire” on page G5 of the Glossary in Carl McDaniel, Jr., & Roger Gates, Marketing Research (John Wiley and Sons, 7th ed. 2007), whereas on page 65 the same authors define pilot studies as “. . . surveys using a limited number of respondents and often employing less rigorous sampling techniques than are employed in large, quantitative studies.”

2. Some research experts define pretests as including both pretests and pilot tests, e.g., “Pretests are small scale rehearsals of the data collection conducted before the main survey. The purpose of a pretest is to evaluate the survey as well as the data collection and respondent selection procedure.” Robert Groves, et al., Survey Methodology 247 (John Wiley & Sons 2004).
A pilot test has a number of potential benefits. For example, it may reveal that the proposed survey methodology will operate effectively because the questions are clear and easily understood by respondents and the method of soliciting participation produces a high yield. Or on the other hand, a pilot test may uncover potential problems such as ambiguity in the questions or difficulties in the sampling frame’s ability to locate qualified respondents—in either case calling for a revision in the survey methodology. Under certain conditions, a pilot test may indicate that a claim or defense is not likely to be supported by a full survey, informing the client and the attorney that dropping or settling the dispute may be warranted.

The purpose of pilot tests is to increase the reliability and validity of the measurement of the state-of-mind constructs the client/attorney, in consultation with the expert, determine to be germane to the matter at hand. Reliability refers to whether the results of the survey could be reproduced by conducting the survey a second time, while validity refers to whether the survey actually measures what it purports to measure. Validity in the context of Lanham Act surveys is particularly concerned with whether the results are both unbiased and unambiguous as to conclusions reached about the state-of-mind intended to be measured.

The fact that properly conducted pilot work was conducted should increase the court’s perception of the care the expert took in ensuring that the survey was designed to be as reliable and valid as possible. Pilot work is consistent with the scientific perspective that the Supreme Court expressed in Daubert because its purpose is to increase the quality of survey evidence.

Naturally the question of whether such efforts were conducted objectively or for a “darker purpose” is a proper focus for cross-examination. Nevertheless, courts on some occasions should be suspicious of a survey where no pilot test or preliminary work by any other name was conducted. Pilot work, properly conducted, should be viewed as evidence of scientific objectivity rather than creative conniving.

As of November 20, 2011, a review of Lanham Act case law identified 25 cases in which the term “pilot” was used in describing such preliminary work, six in which the term “pretest” was used, and one in which both “pretest” and “pilot” were used. In most cases, the court either appeared to find the results of the pilot work to have some value as empirical evidence or did not make specific comment on its value one way or the other. In a few cases, the pilot or pretest results were evaluated by the court as


if the results had been offered as empirical evidence (i.e., as the “final survey”) and in that regard were often not considered valuable because of one or another infirmity, usually either because of small sample size or substantive defect.6 In one case, a pilot was excluded for legal procedural reasons.7 Such criticisms were not directed at the fact that the pilot or pretest was preliminary work; rather, the criticisms were of quite the same sort as typically leveled against a final survey that had been offered in evidence. In some cases a preliminary inquiry that might ordinarily be referred to as a pilot or pretest was treated as if it were the “final” survey.8 Court vocabulary is inconsistent.

From this perspective, it is clear that the court in Lanham Act surveys does not generally view pilot work from a traditional scientific perspective. In scientific pursuits, a properly conducted pilot test would not be evaluated in terms of the evidentiary value of its results in shedding light on the instant issue. Rather, it would be viewed as helping to clarify survey questions or procedures in order to increase the likelihood that the survey would provide more reliable and valid results.9 However the court may choose to view the conduct of such preliminary work, the expert must be able to explain why there was a need for a pilot, what the results of that effort were, and what changes, if any, were made in the final survey design and why. In general, if circumstances dictate the need for a pilot test, it is best for the testifying expert—rather than another—to conduct that study.10 The scientific value of the survey is what the expert strives to improve by conducting pilot work and therefore the expert should not be defensive about having conducted such work.


7. An interesting example is Lisa Frank, et al. v. Impact International, et al., 799 F. Supp. 980 (D. Ariz. 1992), in which a pilot study was reported as having been conducted but was regarded as having “limited probative value” not because it was a pilot test per se, but because it had other infirmities.


9. This was the case in Jellibeans, Inc. v. Skating Clubs of Georgia, Inc., 716 F.2d 833 (11th Cir. 1983), in which the court noted that the fact that the questionnaire was not pretested was an error, and a pretest could have been conducted “... to determine if the questions were confusing.” Also, see Blumenfeld Dev. Corp. v. Carnival Cruise Lines, Inc., 669 F. Supp. 1297 (E.D. Pa. 1987), and 1-800 Contacts, Inc. v. Lens. com, Inc., d/b/a/ LENS.com, 2010 U.S. Dist. LEXIS 132948 (D. Utah Dec. 15, 2010).

10. Obvious legal and ethical considerations preclude the pilot test from remaining hidden by the expert who has been informed by its results, whether or not the expert actually conducted the pilot.
This chapter explores the role that pilot work can play in improving the reliability and validity of the final survey. It describes the kinds of pilot work that might be conducted to inform the development of the survey, but it does not deal in detail with the mechanics of how such studies should be conducted. That detail is more aptly addressed in textbooks and other treatises on marketing research methods and questionnaire construction.11

The types of pilot work methodologies to be discussed include (1) assessing various survey modalities (e.g., personal interviewing, telephone, Internet, etc.) as to the quality of the resulting measurements; (2) trying out the questionnaire on target respondents to improve the clarity of wording; (3) making sure the mechanics of administration are correct (e.g., how the product or advertisement will be presented to the respondent, that the questionnaire “skip pattern” makes sense, etc.); and (4) evaluating the ability of the interviewer (if an interviewer is used) to correctly ask questions and record answers.

The Expert and the Attorney: Different Perspectives on the Value of Pilot Tests

The foremost concern of the expert in designing a survey is validity, that is, whether or not the state-of-mind construct sought to be measured by the survey actually results in a measurement of that construct that is unbiased and unambiguous. To determine the validity of a question or measure in the behavioral sciences, measurements are often correlated with (or validated against) an actual behavior, often referred to as the criterion measure, which is taken to be the “true” measure of the construct or a manifestation of that measure. A simple example in consumer behavior is the prediction of a future purchase from a present measurement scale such as the frequently used 10-point “intention-to-buy” question. The respondent indicates the likelihood of buying a particular brand in the future by expressing a quantitative magnitude of “intention-to-buy” between 1 and 10. If the measurement is strongly predictive of, that is, highly correlated with, actual future purchase behavior, the measure is said to be valid.12


12. Note that the terms “valid” and “reliable” are often used incorrectly in legal treatises and court decisions. Reference is often made to the importance of reliability when it is clear that the court meant either validity only or both reliability and validity. Sometimes the word validity does not appear at all with reference to the survey’s quality, while the term reliability does. Justice Blackmun correctly defined the two terms in Daubert v. Merrell Dow Pharm., 509 U.S. 579 (1993), “We note that scientists typically distinguish between ‘validity’ (does the principle support what it purports to show?) and ‘reliability’ (does application of the principle produce consistent results?). Although these two terms refer to really different concepts, the distinction is sometimes muddled, e.g., ‘. . . the difference between accuracy, validity, and reliability may be such that each is distinct from the other by no more than a hen’s kick.’” Bert Black, Unified Theory or Scientific Evidence, 56 FORDHAM L. REV. 595, 599 (1988).
However, the relevant criterion measure is not as straightforward in surveys conducted in Lanham Act settings. How does one establish a behavior or fact to determine whether an advertisement is “actually” false or if the use of a trademark is “actually” confusing? Some actual behaviors may come to mind, e.g., misdirected communications in cases of alleged trademark infringement are often considered to be evidence of actual (beyond a likelihood of) confusion. However, if one were to attempt to use that measure as the criterion measure for whether or not a survey question is valid—that is, to see if the question accurately predicted that criterion measure—one might have to survey thousands of consumers. That would be the case even if a substantial percentage of consumers were actually confused, because in the typical case only a very small percentage would be apt to voice their confusion (e.g., by misdirecting a communication). And in false advertising situations, a consumer might not know that she has been deceived because in the typical case nothing happens to sufficiently negate or contradict fulfillment of the expectations that she believes she has been deceived. She can only tell us how she understood the representation(s) and then a comparison is made between these perceptions and the “truth.” Were the expectations created by the advertisement true given the actual performance characteristics of the product, or were they false?

As a practical matter, a history of court acceptance plays an important role here. When the so-called “Eveready” questions reveal the presence of a likelihood of confusion, we simply assume that the measure is valid in measuring likelihood of confusion, not because the questions have been validated against an external behavioral criterion measure—they haven’t been—but because they have face validity and also because they have been accepted and relied upon by federal courts in dozens and dozens of decisions. Similarly, in surveys evaluating allegations of false advertising, we typically accept the consumer’s perception of an advertisement or other marketing communication as the criterion measure (is their perception true, or not?) rather than searching for a behavioral manifestation of that perception, since that approach has been accepted and relied upon by federal courts.

The attorney focuses less on what the pilot work reveals about the validity of the survey methodology (although she is wise to consider validity in anticipating the ability to present a defensible case in court) and more on whether or not the proposed survey is likely to produce results favorable to her client. Ultimately, of course, the court’s decision is the only relevant criterion measure for determining the survey’s validity. As previously discussed, the expert, on the other hand, is primarily focused

13. E.g., a consumer calls the plaintiff’s company with a complaint about the defendant’s product, thinking the source of the product to be the same.
14. Union Carbide Corp. v. Eveready Inc., 531 F.2d 366 (7th Cir. 1976). Note that the Eveready surveys used somewhat different questions than were used in the past decade or two, and that neither survey conducted in that matter employed the use of a “control group.” Nevertheless it has generally been considered appropriate to refer to questions used with a control group as “Eveready questions.”
15. In this sense, the attorney thinks of a “pilot” as if it were a “TV program pilot,” and is looking at the pilot as an indication of whether the survey will be “successful” if and when it is ultimately conducted.
on what the pilot work reveals about threats to reliability and validity and how to reduce them. The expert has been trained in perception, attitudes, cognition, marketing research, quantitative techniques, and so on, but usually not in the law or in legal strategy. Thus, it should not be surprising that decisions about conducting pilot work, as well as agreement as to which parameters or variables ought to be the subject of that pilot work, are often not easily made. Since the same rationale must apply to the selection of parameters to be addressed in the final survey, any differences between the expert and the attorney must be resolved to the satisfaction of both.

An Example of the Value of Pilot Work

Suppose that in a false advertising matter, the initial interpretation offered by the attorney as to what state-of-mind the survey should address is that the advertisement is deceptive if it communicates that a product is “low calorie.” Further, the attorney and expert are informed by prior case law that open-ended questions are generally preferred to closed-ended questions by most tribunals. Therefore, the expert might begin by drafting a questionnaire that starts with a “filter” question, “Did you get any understanding from the commercial as to the calorie content of the product?” If the answer is “yes,” then the next question might be, “What, if anything, did you understand the advertisement to have communicated to you about the calorie content of the product?” Suppose that asking this question evoked responses such as “it doesn’t have many,” “it has less than other brands on the market,” and “it has zero calories,” among others.

Which of these answers evidences deception? Among the three listed answers only “it has zero calories” literally evidences deception if the product in fact contains some calories. The other two answers would likely be considered indeterminate. “It doesn’t have many” would indicate the product is perceived to have some nonzero amount of calories, but not necessarily a “low” amount. Similarly, “it has less than other brands on the market” does not necessarily mean that “low calorie” was perceived. To the extent that the question evokes ambiguous answers, an attempt to improve the question’s validity in measuring deception could and should be undertaken.

If this ambiguous outcome were discovered in pilot work, there would be an opportunity to modify the question to increase the likelihood of obtaining a clearer measure, but if found only after the entire survey was conducted, these ambiguous answers would have to “count against” the plaintiff’s position (the number of ambiguous/unclassifiable answers would remain in the denominator in calculating the percentage who were deceived). Improving the question by making it less ambiguous could be in the plaintiff’s interest in light of the plaintiff’s burden of proof, but it could also be in the defendant’s interest if a less ambiguous set of responses would reveal that consumers did not take away a deceptive message. Changing the question should not be viewed as an effort to “cook” the results provided that the question remains unbiased after being modified.

A pilot test adds a dose of reality to the client’s perspective on the merits of his claim, often affecting decisions to proceed or settle.
The expert might take these pilot answers to indicate the need for a more focused yet open-ended question. For example, “What, if anything, did you understand the advertisement to have communicated to you about whether or not the product is low calorie?” One might suppose that this question would evoke less ambiguous answers. However, if answers to the revised question included, “it is not high calorie” or “other brands have more calories” or “it has very few if any calories,” how would one classify these answers as to whether or not these consumers perceive a “low calorie” representation? Very likely the first two answers are still ambiguous, while the third might reasonably be coded as evidencing a perception of “low calorie.” Skillful probing would clearly be required to resolve how the response should be classified.

Ultimately, the expert must consider whether or not, even with careful probing, a classifiable answer is likely to be obtained. Sometimes the consumer’s own construction of an answer to a question does not permit the unambiguous classification of that response as it relates to the legally relevant question. In such a case, the expert might recommend a closed-ended question that would follow one or more of the open-ended questions. An example might be, “If you have an opinion, did the advertisement communicate to you that the product is high in calories or that it is low in calories or something else about the calorie content or nothing about the calorie content?”

The need for a parallel closed-ended control cell or question would be critical in order to evaluate the meaning of the answers. Although a control question or other control procedure is always necessary if a causal issue is being addressed, the need for a control to account for and eliminate spurious causes of answers is greater in the case of a closed-ended than an open-ended question.

Clearly, one could go around the pilot test circle forever. There is probably always some room for improvement in reliability and/or validity. However, at some point the reliability and validity gains are outweighed by costs (in money, in time, and in “suspicion” terms), and pilot work must, therefore, be terminated. Therefore, the initial pilot work should address the concerns most likely to impact reliability and validity. The primary concern must be increasing the odds that the survey will not be rejected on a motion in limine and that, if admitted, it will likely be given substantial weight.16

**TYPICAL RESEARCH PARAMETERS STUDIED IN PILOT TESTS**

The researcher is assumed to be one who has been trained in and is experienced with the design and conduct of a survey in general and for Lanham Act purposes specifically. This training and experience will guide the expert in designing pilot efforts. For example, in writing the initial draft of the questionnaire and in specifying the

---

16. In some cases, the control cell or control questions could yield sufficient evidence to enable a firm basis for interpreting the results of the survey even in the presence of methodological or questionnaire parameters that could have been improved in validity and/or reliability.
overall survey methodology, it is reasonable to assume that the expert’s skill will have
minimized the need for unnecessary pilot work. But unless that prior experience or other
source of skill has focused on the same subject matter (products or services), and on the
same universe (e.g., demographics or another parameter of the relevant universe), and
using the same measurement modality, among other considerations, it is likely that there
will yet be some parameters that might benefit from conducting a pilot test.

There may be interaction among many, if not most, of the parameters (variables
or questions) in that if one question or procedure is modified as the result of a pilot,
another may need to be modified as well. But for the purposes of simplification,
parameters will be discussed as if independent of one another. Questionnaire
parameters are discussed in the next section, followed by two sections in which
methodological variables are identified.

The Questionnaire

Clarity

Are the questions stated as clearly as possible? Is unnecessary complexity avoided?
Does a question force the respondent to assume something is true which is not or
cannot be true for that respondent in order to answer the question? Are the questions
understood in the way the researcher intended and, to the extent possible, do the
questions mean the same thing to all respondents?

Each respondent’s understanding of each survey question will be unique—at least
to some extent—reflecting differences in cognitive skills, education, life experiences,
and so on. Within reason, the goal should be to minimize the amount of error (error
variance) resulting from the lack of uniformity with which the question is understood.

For example, does the question, “How much do you weigh?” assume weight
with or without shoes or apparel? And when you ask respondents about their favorite
breakfast cereal, do you mean at the present time (e.g., on a December day in a cold
climate) or on the average day throughout the year? When you say “favorite” do you
mean what cereal is eaten most often or what cereal is their favorite no matter how
often it is consumed throughout the year? The skill of the researcher in phrasing the
question from the outset should prevent many of these potential ambiguities, but in
some cases the question will miss the mark for reasons that are not obvious. A pilot
test or the kind of talk-through subsequently discussed may reveal hidden ambiguities.

Vocabulary

Is the question tailored to the likely comprehension skills of respondents in the relevant
population? Can a question simply direct the respondent to “rank order your top three
choices” without further explanation or instruction? The researcher’s vocabulary and

---

17. For example, “Suppose you see this store sign as you are driving down the street . . .” assumes the
respondent drives, and “Assume you are in a grocery store shopping for . . .” assumes the respondent
shops in a grocery store.

comprehension skills may not be the same as those of the typical respondent in the survey. If the respondent population is very narrow on the one hand (e.g., children, university professors, etc.) or if it is very diverse on the other (e.g., the total U.S. household population), piloting question comprehensibility is more likely to be important.

**Probing**

Has the respondent’s state-of-mind been reasonably exhausted with respect to what the question seeks to evoke, or would further relevant information be obtained through more or different probing (without browbeating)? To the extent that the respondent’s state-of-mind is incompletely measured, the results are less valid and/or reliable than they would otherwise be. A respondent typically will give only one or two thoughts or ideas in answering a survey question. The respondent may have additional thoughts in mind but will often assume that a brief answer is sufficient. However, unspoken thoughts may be as important or more important than the things initially said. The pilot may reveal the need for more or different kinds of probing beyond the typical “Why do you say that?” or “Anything else you can tell me about that?” or “What do you mean by [use only respondent’s words]?” What may be thought to be “self-evident” and therefore not as likely to be included in the respondent’s answer may be the very thing the researcher sought to uncover with the question!

**Skip Pattern/Flow**

Is the skip pattern being followed correctly? Especially when the questionnaire is self-administered (e.g., as in a mail survey), it is important to be confident that the respondent proceeds through the questionnaire in the order intended and in the same order as other respondents, and that the correct “‘skip’” occurs based on the respondent’s answer. Sometimes it may be permissible for the respondent to go back to review and/or change a prior answer, sometimes not. In a personal, phone, or mall intercept survey that does not use computer-assisted interviewing that automatically controls question order and skip patterns, the interviewer may need to be monitored to ensure that respondents are being moved correctly through the questions.

Often one or more of the questionnaire issues can be addressed through the use of “cognitive interviewing” and similar techniques by other names. Building on theory and empirical findings in cognitive psychology as to how respondents perceive and understand survey questions, respondents with the guidance of a skilled interviewer talk through their thinking processes in comprehending and responding to each question either as they are going through the questionnaire for the first time.

---

or retrospectively after completing the interview. As a result of this “talk through,”
modifications may be made in the kind and extent of probing that is optimal. Probes
are sometimes scripted (e.g., “And why do you say that?”) and sometimes unscripted
(i.e., at the discretion of the interviewer).

Unscripted instructions to probe (e.g., an instruction such as “probe for clarity”) allow the interviewer to probe wherever and whenever increased clarity and/or completeness is thought to be a potential outcome of the probe. Interviewers may observe various cues, including behavioral and/or facial expressions. For example, they may receive specific requests for clarification (e.g., “Does my college sophomore son living at school count as a member of my household?”), detect hesitation in answering a question, observe a disconcerted facial expression, note irritation manifested in a response, or see other signs that a problem might exist. Interviewers as well as respondents may indicate possible problem areas when an incorrect skip occurs, or a question is misread, or other such irregularities are observed. Opportunities for questionnaire improvement may result based on what the respondent or interviewer does or does not do. Video recording or additional monitoring of practice interviews on some occasions is helpful.

**Universe Specification, Sampling Issues, and Incentives**

Will the planned relevant universe description and sampling frame (the means by which a sample will be drawn from that universe) succeed by identifying a sufficient number and appropriate distribution of qualified respondents to conduct a reliable and valid survey, no matter the survey administration modality? A pilot should provide evidence as to whether it is likely that the survey methodology will succeed in achieving the desired sample size and in obtaining a sample that meets other quality considerations (e.g., is an association membership list sufficiently current?). It will also provide additional bases for estimating survey cost. Pilot work may reveal that an initially specified sampling frame (e.g., Yellow Pages) will not result in a sufficient number of interviews to meet protocol specifications. Additional or alternative sampling procedures may be required. Each sampling procedure will have its own costs, and cost variability from one to another procedure may be substantial. Thus, for budgeting, if for no other reasons, a preliminary evaluation is usually worthwhile.

Aside from cost considerations per se, very often there are alternative sampling frames that might serve as a starting place. Each is likely to have its own strengths and weaknesses. A Dunn & Bradstreet or other such directory or listing organization has its own framework for structuring respondents’ job titles and industry affiliations. The way it does will likely be different from the way a trade organization directory identifies members. Can these lists be combined in some way? What is the extent of overlap in these lists? Should a sampling frame be utilized that has already determined (by a respondent’s membership in an organization, for example) that the respondent is

---

licensed or certified in a particular trade? It may be that the currency and coverage of the list is sufficiently old and/or incomplete that it would be better to phone persons who are listed in more current general phone directories and then rely on a phone screen to determine whether the respondent’s qualifications fit the criteria.

Internet survey organizations usually differ from one another in the methodology used to acquire panel members. Hence, the number of panel members who may need to be contacted to achieve the target sample size may vary from one to another Internet panel organization. Even were there no differences among panels of the various Internet survey organizations, there would remain a question as to how many survey participation invitations would need to be disseminated to obtain enough qualified respondents who are willing to participate given the nature and magnitude of the incentive offered for their participation, the length of the survey, and other characteristics (e.g., is the survey about personal products about which some unknown number of possible respondents will choose not to participate)? Many Internet survey organizations make available a free or low-cost omnibus (multiple client) survey that is often conducted weekly or monthly and that is specifically addressed to this qualification or incidence issue. One may also discover that some definitions of “relevant universe” are not likely to yield a sufficient number of interviews to proceed. For example, yield may be low for respondents suffering from a rare disease, homemakers who add unusual condiments or ingredients to food dishes, and so on.

Optimal incentive (token compensation or “thank you” payment for survey participation) magnitude will also affect cooperation/response rate, and a split-run pilot test can be conducted to determine which of several different incentive magnitudes would be the most productive in obtaining survey respondents’ cooperation. It is not necessarily the case, of course, that any incentive is better than none at all, or that a higher incentive always achieves a higher cooperation rate. However, a pilot provides a basis for cost-benefit determinations.

Mechanical, Administrative, and Logistical Considerations

Choosing the Survey Modality

The nature of the posited state-of-mind in which deception or confusion occurs usually dictates the survey modality (mall intercept, telephone, Internet, etc.) that is optimal, but some alternatives may exist. For example, if alleged confusion is thought to occur as the shopper makes a product selection decision at a supermarket shelf, then a likely modality is mall intercept, with the respondent being exposed to a simulated store shelf. Alternatively, perhaps the interview could actually be done in a supermarket.21 At the planning stage, two (or more) possible modalities may compete in the mind of the researcher. A pilot test could consist of a small sample test using two different modalities so that additional reasons for choosing one over the other may become apparent. The modality choice will necessarily interact with the nature of the

21. See Kumar, supra note 11 at 219–220.
stimulus to be tested. A 30-minute infomercial as compared to a one-minute television commercial as the focus of a false advertising survey will significantly influence the choice of survey modality. Similarly, questionnaire length will interact with the optimality of alternative survey modalities.

**Interviewer Training**

In telephone or personal surveys, the interviewer is another variable in the survey equation that can have a great impact on survey quality. Just as question phrasing may not be clear to the respondent, some instructional content or procedural detail may not be clear to the interviewer. In some cases, pilot work may reveal that some interviewers erred in how or when probing was conducted, when to present one stimulus rather than another, or whether the stimulus properly remained in view (or not). The need for a change in, or clarification of, interviewer instructions may become apparent.

**Practice Interviews**

Interviewers may benefit from verbalizing their thoughts about survey questions while practicing interviews with one another so that the need for changes in administration instructions (or question content) becomes apparent. Whether the survey uses open- or closed-ended questions, the researcher may benefit from reviewing or listening in to the first several practice interviews to be sure that there are no obvious administrative problems. The extent of the training that will be necessary is sometimes not clear until one conducts practice interviews. If the interviews are being conducted properly, the formal survey can begin. Otherwise, changes will be made and the survey restarted.

**Recording Answers**

In telephone or personal (e.g., mall intercept) interviewing there may be occasions when survey answers should be tape-recorded in order to capture the necessary level of detail. When administering surveys where many open-ended questions are being asked, and especially when the relevant universe is technical or professional where specialized vocabulary is used, the interviewer may not be able to write fast enough, yet legibly, to capture all of the important points (or verbatim content, in some cases) of the respondent’s answer. One cannot know for certain the answer to questions about likely interviewer performance without actually conducting some interviews to evaluate the outcome.

**HOW LIKELY IS IT THAT PILOT WORK WILL BE HELPFUL?**

**Piloting Usually Provides at Least Some Helpful Information**

Experienced researchers recognize that preliminary work is likely to inform the research process. As previously discussed, attorneys are more likely to look at the results of pilot work for other reasons, that is, as an indication of whether the proposed
survey is likely to produce favorable results. A survey that produces favorable results, however, has the potential to aid the client’s interest only if the survey is admitted into evidence and given some weight. Therefore, it follows that there is usually some value in a pilot test, for the attorney and client as well as for the expert.

Although prior experience is important, to the extent that the circumstances of the instant survey are different from prior surveys, experience may not be a sufficient guide. There is always something different about the circumstances and variables involved in one survey situation compared to another. If what seem to be unimportant differences turn out to have a substantial impact on survey outcome it goes without saying that it is better to be safe than sorry.

Circumstances in Which Pilot Work Is More Likely to Be Helpful

The more the instant matter differs in one or more respects from matters examined in prior surveys, the greater the potential value in conducting pilot work. In some instances, the state-of-mind construct the survey is being designed to measure may be ambiguous and/or the kind of “confusion” or “deception” alleged may be atypical (as compared with case literature examples).

For example, a plaintiff alleges that the appearance of a newly constructed retaining wall infringes the plaintiff’s trade dress, claiming that consumers passing by the constructed wall in a car or bus or other vehicle, whether as driver or passenger, would likely be confused as to the source.22 It is alleged that they would think the wall they are looking at is made from Brand X when it is in fact made from Brand Y. Can the survey be reliably administered by interviewing respondents as they actually drive by or are driven by the wall in a vehicle? Or should the drive-by be simulated in a visual prepared for mall intercept or online Internet administration in order to maintain consistency among all viewing experiences? Are there other possible designs that could be evaluated?

Circumstances When a Pilot Test Is Less Likely to Be Helpful

In some circumstances, there is likely to be limited value in conducting a pilot test. For example, sometimes a replication (with modification) of the other party’s survey is considered to be the most useful kind of survey to conduct.23 The expert believes that the only thing seriously wrong with the other party’s survey was that the stimulus used as the control was incorrect or that the relevant universe was defined incorrectly. In each of these two circumstances, the researcher’s recommended survey design may well be to repeat exactly the other party’s survey except for a change in the control stimulus or a change in definition of the relevant universe. It would generally be

undesirable for changes to be made in other parameters of the survey, even if some of them would not ordinarily have been chosen by the expert if he or she had been asked to conduct the survey from the outset. If the other side has deemed a survey design to be valid and reliable, then obviously adopting the same design, but with a change in one or two key parameters, lessens the potential criticism that could be directed at the (modified) survey replication.

Sometimes a prior reported decision that has been relied upon by the same trier of fact involves a survey methodology that the attorney suggests should be followed to increase the likelihood that the court will rely on another survey using the same methodology. Provided the expert does not so fundamentally disagree with the design of the questionnaire he or she is being asked to implement, the expert is ordinarily guided by that judicial history. Of course, the expert must be prepared to defend the approach taken. Would the expert have chosen that particular design if it were up to him or her?24

PILOT TEST IMPLEMENTATION CONSIDERATIONS

Pilot Survey Sample Size
In the typical case, a pilot will consist of between 10 and 50 respondents, but many factors unique to the particular questionnaire and overall survey execution may modify that sample size. For example, a pilot test of a long questionnaire usually requires more respondents than one that is short; a survey with many open-ended questions requires more respondents than one with mostly closed-ended questions. If the pilot test is to include methodological and administrative issues, then it requires even more respondents.

How Many Parameters Can Be Evaluated at the Same Time?
If one intends to compare one research design or survey question versus another, in order to be sure that differences in pilot test results can be attributed to a one-versus-another variable, it is ideally the case that only one parameter be evaluated at a time. Otherwise it may not be clear from the results whether one or the other (or both) changes caused a change in the results.

And, since changes in any one parameter are likely to cause the need for modification of some sort in one or more other parameters, it is usually best to study only one change at a time, when possible. For example, changing an open-ended to a closed-ended question will likely require a change in what are determined to be necessary

24. However, even where such guidance is given by the attorney, a reported decision may provide only a superficial description of what was actually done in the prior survey. Assuming that the actual testimony/report of the expert who conducted the prior survey is not available, the expert may be well advised to pilot one or more interpretations of what might have been done to aid in deciding what would be the best approach in addressing the instant matter.
“control” questions. Parameters that appear to be most central to the reliability and validity of the survey’s outcome have priority.

Pilot Study Timing
The pilot test must be conducted close enough in time to the final survey so that it can, in fact, be guiding. There are many reasons this is true, including primarily (1) one cannot guarantee that the other party or that prospective survey respondents in the final survey do not become aware of the fact of a pilot test, which could contaminate the final survey, and (2) one or both parties may change advertising, packaging, or trade dress of the particular products of interest, changes that could invalidate what would otherwise have been guidance from the pilot.

How Many Pilot Tests?
There is conceptually no end to the number of pilot tests that could be useful: something might be gained each time. But there are diminishing returns in value because total survey costs would increase while the reliability and validity gains would be less and less. One or two pilot tests are ordinarily sufficient.

CONCLUDING THOUGHTS
Good pilot work seeks to make clear what is unclear, or at least to improve the clarity and relevance of the survey’s end result. This attempt to improve the reliability and validity of a survey is not completely separable from the methodological advice provided in other chapters in this book. For example, much of this chapter deals with improving question structure. To a large degree the expert should know, on the basis of training and experience, that phrasing a question this way rather than that would improve the clarity of the response. This would be true regardless of the subject of the survey’s investigation. A pilot test that reveals that, indeed, the questions as structured do not require improvement wastes time and resources that could have been spent on piloting issues that were less certain. Thus, open-ended versus closed-ended questions, the extent of probing, and the methods by which responses are probed, among other topics, are not ordinarily what would be addressed in a pilot test, except (as in the examples given) when there is uncertainty about which levels and kinds of probing are optimal in the special circumstances of the relevant state-of-mind at issue.

The question for the court should be: “Was the pilot test conducted fairly to further the interests of reliability and validity?” Did the pilot test have “evidentiary reliability”? Or, on the other hand, “Was it conducted simply to improve the value of the results for the client?” This does not mean that a pilot test cannot result in changes that do, in fact, benefit the client’s position. In fact, improving the quality of the survey should remove potential obstacles from a valid claim or defense. The question is

25. Blackmun in Daubert, supra note 12.
whether the pilot test was somehow rigged or slanted by design to achieve that outcome. Understandably pilot tests will add the opportunity for even more “battles of the experts” at a time when some courts are already dubious of the objectivity and honesty of consumer surveys and of the experts who testify about them.

Embracing the value and acceptability of pilot testing will likely turn on whether the court and other tribunals dealing with similar issues (e.g., FTC, NAD, TTAB, etc.) adopt a more scientific perspective. Even though the Daubert standards were a move forward, it also follows that adoption of this perspective in the language of the Manual for Complex Litigation, in tomes on surveys (e.g., McCarthy on Trademarks and Unfair Competition), in learned articles published in the legal and psychological/marketing research literature, and so on, would be helpful.

It will be recalled that it has not been so many years since the concept of a control group found its important place in consumer survey methodology. And many attorneys still flinch at the use of the word “experiment” because they are concerned that a survey that is also (or rather) an experiment (these days, the great majority of Lanham Act consumer surveys) would be perceived as “less certain” than a survey that is not also an experiment. Fortunately, most courts understand the difference between “experiment” as in “this is untried or tentative but we’ll see what happens” versus “an experimental design allows one to more clearly assess cause and effect.”

Change comes to all disciplines, to some more slowly than others.