This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30049010

Every reasonable effort has been made to ensure that permission has been obtained for items included in Deakin Research Online. If you believe that your rights have been infringed by this repository, please contact drosupport@deakin.edu.au

Copyright: 2012, International Trademark Association
COLOR, CHAMPAGNE, AND TRADEMARK SECONDARY MEANING SURVEYS: DEVILISH DETAIL

By David H.B. Bednall, Phillip Gendall, Janet Hoek, and Stephen Downes

I. INTRODUCTION

Survey evidence is often adduced in commercial law proceedings, most commonly in trademark cases. Although previously rejected as hearsay or of little material value, survey evidence is now accepted in many jurisdictions in the United States and elsewhere. In the trademark domain, surveys have focused on four areas: passing off (or consumer deception), evidencing whether a name or symbol is generic, trademark dilution, and testing whether a word, symbol, or color acts as a badge of origin for a brand, thus supporting its registration as a trademark. This article focuses on the fourth area and argues that past commentaries on survey designs in this area, by failing to
consider how survey design elements necessarily interact, do not fully describe the dilemmas faced by researchers in producing implementable and defensible designs. These "devilish details" can be the undoing of any survey design.

Judicial authorities and the courts themselves have issued guidelines as to how surveys should be conducted and described to the court. The following guidelines were listed by the Federal Judicial Center in 2004:

The sampling methods used must conform to generally recognized statistical standards. Relevant factors include whether

- the population was properly chosen and defined;
- the sample chosen was representative of that population;
- the data gathered were accurately reported; and
- the data were analyzed in accordance with accepted statistical principles.

In addition, in assessing the validity of a survey, the judge should take into account the following factors:

- whether the questions asked were clear and not leading;
- whether the survey was conducted by qualified persons following proper interview procedures; and
- whether the process was conducted so as to ensure objectivity (e.g., determine if the survey was conducted in anticipation of litigation and by persons connected with the parties or counsel or by persons aware of its purpose in the litigation).

Courts also have available to them the Reference Manual on Scientific Evidence, which contains a technical guide to the use of surveys by the courts. Additionally, McCarthy has extensively reviewed the use of surveys in trademark litigation. This analysis, along with reviews of faults with previous studies, has provided additional guidance for future studies. The key areas may be summarized as follows:

---

11. 6 McCarthy, supra note 1, § 32:158 to § 32:163, and § 32:179 to § 32:196.
12. Fred W. Morgan, Judicial Standards for Survey Research: An Update and Guidelines, 54 J. Mktg. 89 (1990); Richard Gough, Gough on Survey Evidence in Australia,
1. Nonresponse should not bias the survey results;\textsuperscript{13}
2. Questions must be clear, unambiguous, and non-leading;\textsuperscript{14}
3. Filter questions must be used to question only interviewees who are knowledgeable;\textsuperscript{16}
4. Questions must be relevant to the issues and unambiguous;\textsuperscript{16}
5. The use of control questions\textsuperscript{17} or a control group must be considered;\textsuperscript{18}
6. The survey’s methodology must approximate market conditions;\textsuperscript{19}
7. The stimuli shown should closely resemble their actual deployment;\textsuperscript{20}
8. Interviewers must be properly selected and trained and the survey must be properly administered;\textsuperscript{21} and


14. 6 McCarthy, supra note 1, § 32:172.
15. Rivera, supra note 12, at 681. This paraphrases his argument. An issue could be taken with his comment here that “guessing . . . affects the reliability of the survey by increasing the survey’s random error.” Guessing may be systematic, for example, by interviewees nominating the most common brand, rather than being random. However, it is agreed that interviewees must be in the target market. Depending on how the survey recruitment is framed (see below), this should be people who are or who have been wine or champagne purchasers.
17. Rivera, supra note 12, at 666-668. This paraphrases his recommendation. It is arguable whether they are actually control “questions” or control conditions, as is discussed below. Essentially the issue is whether responses to a stimulus with the color are different from responses to the same question asked of a stimulus using the control color. Control questions are more likely to be useful in litigation involving consumer confusion, where confusion between the two brands in contention should be set in the context of control questions about possible confusion between other brands in the category.
18. 6 McCarthy, supra note 1, § 32:187. While this was discussed in the context of confusion, the same issue arises in secondary meaning surveys. The use of two stimuli that are identical except for color enables the influence of the color on responses to be directly assessed. See also Diamond, supra note 10, at 398.
19. 6 McCarthy, supra note 1, § 32-163, at 32-335.
20. 6 McCarthy, supra note 1, § 32:171, at 32-358 to 32-359. This summarizes his argument about the need for realism in what is shown to interviewees. This is somewhat in conflict with a view that is irrelevant to show . . . interviewees the whole product, which included nonprotectable, functional features.” In order to disentangle consumer response stimulus versions with and without the “nonprotectable, functional features” could be used. In this way, possible arguments about the lack of realism of the stimulus materials could be overcome.
9. Interviewers must be monitored.  

With the exception of questionable descriptions of filters and survey controls, these guidelines are sensible and unobjectionable. However, criteria such as these commonly fail to show how the various elements of a survey interact, such as the sampling with the data collection method, and do not always specify how the research task or topic should be framed for presentation to interviewees. They may also lack guidance on crucial implementation details. As will be argued, previous trademark studies also suffer from the limitation of testing secondary meaning through measuring various kinds of association, not identification.  

A specific survey design is proposed that addresses these four issues. Paradoxically, paying attention to the "devil in the detail" of this specific design raises fundamental issues for all such surveys. The proposed survey seeks to examine color in connection with champagne, that is, whether a particular color, namely orange, acts as a trademark for the VEUVE CLICQUOT brand of champagne. This color is not registered in the United States, and though it has attained registration in two jurisdictions, no survey evidence was adduced in either.  

This article first outlines the requirements for trademarking color and describes the VEUVE CLICQUOT brand. It then describes possible designs for testing whether the color orange serves as a trademark. For each element of the survey, the design dilemmas that must be resolved prior to implementation are described, and specific solutions to these design dilemmas are suggested. As a final contribution, the article looks at arguments that may be raised against the design elements and suggests ways that these arguments may be anticipated and minimized by using a specific design logic. In this way, this article will help expert researchers produce specific, defensible designs that will be of use to courts and tribunals.

---

22. 6 McCarthy, supra note 1, § 32-170, at 32-355 to 32-356.  

23. It is recognized that in most jurisdictions, the legal test is association. As will be argued below in terms of survey evidence, asking about association rather than identification may produce ambiguous evidence.  


II. TRADEMARKING COLOR

Once they have established a brand, companies seek to build and then protect their brand equity.\(^{26}\) Registering a trademark for their brand, product, or company is one form that protection takes. “A trademark is a word, phrase, symbol or design, or a combination thereof, that identifies and distinguishes the source of the goods of one party from those of others.”\(^{27}\) The registration of color trademarks remains legally contentious, but since the landmark Qualitytex v. Jacobson case,\(^ {28} \) courts in the United States, Europe, and elsewhere\(^ {29} \) have allowed the registration of a single color as a trademark.\(^ {30} \) Such cases are hard to win, even when courts in other jurisdictions have allowed registration.\(^ {31} \)

In making a case for a symbol to be registered as a trademark in the United States, the owner needs supporting evidence. The main type of evidence, which is the focus of this article, is that the symbol is shown to be distinctive in use and to have acquired a secondary meaning.\(^ {32} \) In part, this can be demonstrated by the history, use, and exclusivity of the symbol and by extent of sales and advertising of the marked product as well as direct consumer testimony relating to the product.\(^ {33} \) Ultimately, however, direct consumer testimony is limited in scope, meaning that systematic

\(26\) Kevin L. Keller, Strategic Brand Management: Building, Measuring and Managing Brand Equity (3d ed. 2008).


\(30\) As Vana (see supra note 28) observes, specific colors need to be considered in the context of possible shade confusion. While the Pantone Matching System (PMS) specifies specific colors (see http://www.pantone.com), the colors historically used by some brands predate the widespread use of PMS color specifications.

\(31\) Gulenay Rusen, Registrability of colors and smells as a trademark, 2 Int'l J. Intell. Prop. Mgmt. 75 (2009). This is not to imply that trademark laws, markets, or precedents in various jurisdictions are the same, or that one country must take account of outcomes in other countries. Rather it is to say that with global markets, similar marketing of a mark and similar consumer responses to it may lead to similar arguments and evidence for registration.

\(32\) Swann, supra note 3.

survey evidence is commonly sought. This evidence serves to test whether buyers or consumers distinguish the brand on the basis of the trademarked element. A more liberal definition of secondary meaning says the "mark or dress identifies one product," while a moderate definition says the "mark or dress identifies the product of one, perhaps anonymous company" and a conservative one that the "Mark or dress identifies one company." In applying for registration on this basis, the applicant is asserting that at the time of the application, the symbol was already distinctive for consumers.

III. THE VEUVE CLICQUOT PONSARDIN COMPANY AND BRAND

Veuve Clicquot Ponsardin traces its origins to the eighteenth century. The company, which is based at Reims in the Champagne region of France, is the maker of Brut Champagne, its premium brand. Brut, described by Veuve Clicquot Ponsardin as "Yellow Label," is the company's most famous product and has been in production for over 120 years. Reportedly, VEUVE CLICQUOT PONSARDIN is the second largest selling champagne brand in the United States, a relevant fact when considering the brand's inclusion in a trademark survey. The survey will test the color itself, not the word descriptor of the color. A layman might describe it as orange. The product's bottle, made of a traditional dark green glass, has a colored label on its face, a gold foil enclosing the neck and top, and an orange-colored neck label at the base of the foil. The main label, which is rectangular, is entirely orange on the background with the brand name ("Veuve Clicquot Ponsardin"), the words "Champagne," "Maison Fonèée en 1772," and "Reims France," in black, a stylized logo in black, and a signature in red. The label has a rectangular muted black border near the edge. The product is also marketed in an orange box. A reproduction of its standard bottle with label and packaging is shown below. The company has taken many legal steps to protect its branding by registering various trademarks and by litigation.

35. Evans & Gunn, supra note 8, at 26.
40. Registrar of Trade Marks, supra note 25.
IV. DESIGNING A SURVEY TO TEST THE DISTINCTIVENESS OF THE VEUVE CLICQUOT ORANGE COLOR

Many surveys designed to test the legal distinctiveness of color have been conducted in the United States and elsewhere. The role of the survey is not to decide the legal issues, but to complement other evidence. It is likely that such surveys will be conducted mostly for dominant brands or those that are major players in their class of goods. This is particularly likely in fast-moving consumer goods markets or where the products are high-status brands in a market niche and where there is a history of the development of the symbol or mark in conjunction with the brand. An example of this would be Mars Confectionery and its cat food brand, Whiskas.

Frequently, registration of a color trademark is opposed by one or more parties. Normally, survey research is commissioned by the party seeking registration in the belief that the research will support its claims. Any weakness in design or reporting will...
quickly be discerned by the opposing parties. While some faults in a survey do not necessarily prevent its being considered by a court, its value will be greatly diminished. Typically, opponents engage well-known and well-published academics or practitioners to critique the survey. Such critiques can be quite unforgiving in the sense that any imperfection in design, no matter how minor, is likely to be discovered and presented as a potentially devastating flaw. Thus, as discussed below, inevitable critiques must be anticipated as part of the design process.

For the purpose of establishing the survey design, consider a hypothetical application made in 1996 for registering the VEUVE CLICQUOT ORANGE as a trademark. It is assumed that the survey will be conducted among consumers in 2012. For each major survey component, a detailed design will be described and design dilemmas identified. Possible criticisms of the design are debated and solutions suggested.

A. Creating a Double-Blind Study

Color trademarks function by taking the consumer from the sight of the color (the stimulus) to the identification of the brand (the response). In order to account for the response in trademark surveys, the effect of the color stimulus must be isolated from the possible effects of other stimuli that are present. For example, if a consumer identifies a bottle of wine as VEUVE CLICQUOT, that may come about because of various features that the consumer is shown, such as the orange color, a logo, a signature on the label or the foil wrapping on the neck of the bottle, or some combination of these and other features. To separate out the effect of the color, it must be tested on its own and possibly in combination with other features. If a color whose registration is sought is never displayed entirely on its own in practice, then at least two versions may need to be tested. For example, if BP Petroleum seeks to register its green as a trademark, but on its pumping stations it is always displayed with yellow lettering, both green on its own and green with yellow combinations may need to be tested. As Jacoby has pointed out, this means that a standard experimental design is required. By systematically varying the elements in the stimulus,  

44. Swann, supra note 3.
45. Rivera, supra note 12; Thornburg, supra note 5. It is probably simpler for a court to admit survey evidence but give it little weight than it is to subscribe to adversarial submissions that it be rejected in its entirety.
47. These designs had their origins in biology and psychology. The word "experimental" does not mean untested in this context; it means the application of scientific research methods to the survey design. See William R. Shadish, Thomas T. Cook & Donald T.
the effects of the color can be separated out from the other elements.

In addition, the fact that VEUVE CLICQUOT may be one of the few champagne brands familiar to consumers may lead them to guess that the stimulus in front of them is VEUVE CLICQUOT. Medical research has taught the lesson that a mere treatment intervention is enough to trigger a placebo effect, irrespective of the physiological efficacy of the treatment. Double-blind experiments, where neither the person administering the medicine, nor the person taking it are aware of which medicine contains the active ingredient and which the placebo, were developed for this reason. In addition, the conditions of presenting and taking the medicine and the placebo should be directly comparable. Because it is likely that the placebo itself will produce a response, the medicine being tested must generate a response significantly bigger than the placebo for its efficacy to be demonstrated. In the case of VEUVE CLICQUOT at least one control color (equivalent to the placebo) must be used in order to estimate the likelihood of association with VEUVE CLICQUOT irrespective of which color is used. Thornburg discusses the need for controls, but not their particular relevance to secondary meaning surveys, which must control for survey participants' tendency to respond with the names of familiar brands.

The analogous double-blind approach for trademark surveys must ensure that no interviewer, field supervisor, or interviewee should be told the purpose of the study or that its design is aimed at testing a particular color. Nor should any of these participants be informed of the identity of the ultimate sponsor, what types of outcomes the sponsor would find helpful, nor which stimuli are the test items and which the control items. In addition to testing the stimulus (color) of interest, one or more neutral (or control) stimuli are required to act as placebos, which provide a baseline measure of the effect of administering the survey. Critics of these surveys may claim that interviewees or interviewers may form a hypothesis about the purpose of the survey. But if care is taken to

---


48. Id. at 507. "An experiment in which both the treatment provider and treatment recipient are unaware of which treatment or control condition is being administered, primarily used in medical clinical trials." Arguably, what is being proposed is a triple-blind trial, as the field supervisors who manage the interviewers should also be unaware of the legal nature of the survey or the parties involved.


50. Diamond supra note 10, at 410-411. The desirability of not telling the interviewers the sponsor of the survey calls for a double-blind procedure to be implemented.

51. Morgan, supra note 12.
keep them from knowing the nature of the research and who would benefit from it, it is difficult to see how such hypothesizing would lead to a systematic bias. The use of the control (placebo) color is a further safeguard.

In all these circumstances, there is a need to compare responses to the test stimulus (VEUVE CLICQUOT ORANGE) with responses to control stimuli that differ only with respect to color. Some stimuli may contain additional cues (such as package shape or size) to disguise further the key stimulus of interest. Further cues may also be relevant for generalizing the results to a range of possible products. If interviewees see the color orange as strongly identifying the VEUVE CLICQUOT brand across a range of bottle shapes or sizes, then the case for registration is far stronger. In experimental design, this also helps address the external generalizability issue—that is, how relevant is the survey design to the real-life behavior of consumers? The more contexts in which the effect is shown, the greater is the confidence in its generalizability. In research terms, this requires the use of a factorial design. Here, the factors (the stimulus elements—color, shape, or other features) are systematically varied so the effects on the identification response by any one element of the stimulus, either alone or in combination, can be assessed.

B. Framing the Context

A crucial issue is how to frame the topic. While the legal nature of the proceedings should not be mentioned, some frame as to the nature of the survey has to be given in order both to recruit relevant survey interviewees and to have them consider issues relevant to the survey. But this poses a dilemma: “Too broad and we get a lot of noise, too narrow and we might be accused of leading or ‘tipping off’ survey respondents.” Two key elements must be controlled in framing the topic of the VEUVE CLICQUOT ORANGE survey. The first is the color. Either mentioning “color” or saying a word like “orange” or showing a color with no other cues (for example, a swatch of the VEUVE CLICQUOT ORANGE color), will signal that the survey is about color. A second issue is whether the product class—“wine” or “champagne” or “sparkling wine”—is explicitly mentioned or shown by use of a package such as a champagne or other wine bottle. The presence or absence of either set of cues gives four possible ways of framing the study both in terms of recruiting people and showing the stimuli to interviewees.

52. Id. at 948. The discussion covers how many controls might be needed. In the case of seeking to establish whether a single stimulus acts as a badge of origin, a single control might be sufficient. However, if it is intended to generalize the findings to a range of product combinations (e.g., BP green with or without yellow writing), then multiple controls would be needed.

53. Gough & Downes, supra note 34, at 8.
1. No Color Cues With No Product Context

The following could be said in order to recruit interviewees: “We are doing a consumer survey. We want to show you a product label and then get your reaction to it.” The problem here is that registration is sought for a particular type of goods, so if the product class is not mentioned, then the sample of people recruited do not necessarily match the target population.\textsuperscript{54} while other people may feel the request is too vague and refuse to participate as a result. There is little that can be done to prevent this within this design. Those who did agree to be interviewed would be shown a randomly rotated list of products and asked to say which ones they had bought or used. If wine was included in the list, they would be asked the survey questions. The others would be screened out. Although the interviewer knows the survey is about wine buyers, the interviewees probably do not. However, those consumers who said they had bought or used wine may still be more likely to give wine-associated responses than those who were not screened in this way. Thus, this screening takes the approach somewhat away from the “no product context” design ideal.

Qualified interviewees who agreed to participate could be shown a flat sample of the color in the form of a label. Putting the label on any other shape, such as a bottle, will provide product category cues, the very bias this approach is designed to avoid. Thus, when shown the label, interviewees could be asked, “Please look at this product label. Does this label identify any particular brands, products, or companies to you, or not?” If the answer is “yes,” then ask, “Which brands, products, or companies?” The form of questioning is open-ended and does not suggest that any particular response is being sought. Although the questioning does not focus on color, because color is the only cue available apart from a nondescript shape, it reveals that the survey is about color. The use of one or more control (placebo) colors will help put the responses into perspective.

However, there are two further problems with this approach. The first is the issue of spreading activation in semantic memory.\textsuperscript{55} Memory consists of associations, some weaker, some stronger, arranged in a semantic network. Some pathways will take the consumer away from the concept of interest. For example, the orange color may lead the person to talk about oranges or orange-flavored chocolates, the country Holland, a cell phone company using the name “Orange” or a juice bar, because they are more

\textsuperscript{54} Diamond discusses the problem with “coverage errors” when the survey sample is drawn from a population wider than the community of interest, though this is “less of a problem” than a defined population or sample frame that excludes relevant people. Diamond, supra note 10, at 378-379.

salient than wine at the time of questioning. For example, interviewees may have just bought oranges, visited the juice bar, or made a call on their Orange cell phone. This may take them away from any possible association with VEUVE CLICQUOT ORANGE, even if they have knowledge of the color used by VEUVE CLICQUOT. In other words, it may lead to an underestimation of the strength of linkage between the orange color used and VEUVE CLICQUOT, because once the consumer has gone down one branch of the semantic network, there may be some distance to the VEUVE CLICQUOT brand. Of course, a follow-up question that then asks about wine or champagne, if the interviewee has not yet mentioned it, would be needed. But again, the response may be biased away from the VEUVE CLICQUOT response because the consumer’s thinking has been diverted elsewhere. It is, therefore, unclear what value the first question has actually contributed, while the use of the follow-up question has framed the survey as being about “orange/wine,” the very thing the non-contextualized design was aiming to avoid.

Next, the question is inherently directed at finding associations—“what comes to mind,” rather than raising the identification issue. It may be argued that if far higher associations are shown for the VEUVE CLICQUOT ORANGE than for a control color, then that is sufficient to demonstrate that it acts as a trademark. It is argued below that this is actually insufficient. A possible variation, therefore, is to have the first question directly about identification, “Please look at this label. Does it identify any particular brands, products, or companies?” While this does not overcome the salience question and frames color as being the issue, it is more likely to yield relevant information. However, as will be discussed below, the word “identify” is also contentious.

An alternative approach is to use a distractor item so that the product context is obscured. Interviewees could first be shown an item such as a chocolate box in a lime color. Chocolates would be one of the categories asked about in the screening. The survey participants would then be asked about the chocolate item shown. Finally, interviewees would be shown another stimulus in the VEUVE CLICQUOT ORANGE or control color. This stimulus would still need to include more features than just the color, such as a wine bottle with a colored label, but the consumer would be presented with a range of packaging, color, and product cues so that there would be no emphasis on any one of them. The only problem is that the distractor item cues might lead consumers down irrelevant semantic pathways, reducing responses to the test stimulus.
2. No Color Cues With Product Context

In recruiting for the survey, people in this scenario would be told the product category. As will be shown, only a minority of Americans buy or drink champagne. A choice has to be made as to the specificity of the product category used. While it would be possible to talk about “sparkling wine and champagne,” there is a danger, possibly remote, that it would trigger well-known brands and the colors associated with them. If that happens, a possible direction of the association triggered is from brand to the color. In contrast, it is the opposite direction that is being tested, namely, whether the color takes the consumer to the brand. The response to the specific cueing may, therefore, influence what is being measured. While control colors would help estimate the response to colors, the possibility is best avoided in order to prevent the brand cueing the color. Thus, a broader category, wine, is proposed. Use of a broader category also corresponds with the reality that in many outlets where wine is sold sparkling wine and champagne are available.

Thus, consumers would be asked if they had ever bought or consumed wine. Those who respond positively are the most relevant people. If this group does not identify VEUVE CLICQUOT on the basis of the VEUVE CLICQUOT ORANGE, the responses of other consumers are basically irrelevant. Once recruited, they would be shown a wine bottle with a particular combination of color and other package cues, thus avoiding an emphasis on color specifically. This approach is relatively simple, but it could be argued that mentioning wine is predisposing the consumer to think of the wines in the category that they know, especially if the package is in the shape of a champagne bottle. However, this prompting is comparable to the situation that consumers face when they walk into a wine shop and see packages in various colors and shapes. Registration is sought for a class of goods, not for all products and services that use the indicia for which a claim is made. In addition, the use of one or more control colors would form a baseline against which responses to the VEUVE CLICQUOT ORANGE could be judged. It may be argued that the systematic variation of package shape (size of bottle, shape of bottle, and variation in label shape) along with various placebo colors, would enable the effect of the particular color to be discerned, since responses to packages that are identical in all respects, except color, would be compared. The use of a distractor item as in the previous approach would be unnecessary, avoiding some survey complexity and cost.

It could be argued that variations such as those in package size, labelling, and packaging are unnecessary in secondary meaning surveys. But what if an opponent to registration were able to show in a survey that there was an identification response
to an ordinary champagne bottle, but not to a magnum or to a champagne box or to a bottle where the label included the minor use of other colors? It is hard to see how registration for a class of goods could be granted in these circumstances.

Essentially, what the survey needs to test is whether or not consumers, when they experience the orange color, use it to identify a particular brand or product, namely, VEUVE CLICQUOT. In other words, a critical test is whether the color serves as a badge of origin for them that is both clear and unambiguous. This approach is consistent with that requirement.

3. Color Cues and No Product Cues

The recruiting of interviewees would proceed in the same way as the first approach. Various ways of presenting color have been suggested, such as using a color wheel or a flat swatch of color. Consumers are then asked to identify what “brands, products, or companies” are associated with each color or are identified by it. The problem is that the consumer is thus forced to focus on color and to respond in terms of color, given that they have been recruited for a survey where they are expected to have some answers. A contrary bias is that their reactions may lead them down semantic pathways that have little or no relationship to wine or champagne or VEUVE CLICQUOT, artificially lowering their degree of response to the product. If alternative colors were associated with other product types, this distraction would be aggravated. Ultimately, whatever the opening preamble, the survey participants would eventually have to be introduced to the wine category or shown a wine bottle. Until the product category is introduced, the researcher cannot be sure that all possible identification responses to the color stimulus have been uncovered. Finally, trademark registration applies to a class of goods, not to goods in general, so the overall relevance of this approach is highly questionable.

4. Color Cues and Product Cues

The process of recruiting interviewees would be the same as for the second alternative. Interviewees would be shown a wine bottle with a colored label and then asked questions about the color. Assuming the wine bottle were generic, responses would focus on color, with the expectation among interviewees that as they were invited to take part in the survey, they should be able to provide answers. The use of a color wheel, where interviewees are asked to say which color was associated with one or more than one

champagne brand, would be a better alternative. Again, this is forcing consumers to focus on color, although at least it has the benefit of seeing whether the color takes the consumer to the brand in question. As described below, some of the secondary meaning question formats used in past cases would explicitly mention color.

5. Conclusions About Framing

These four approaches represent all the possibilities for framing the survey topic in terms of product and color. All have issues. Prima facie, the first two alternatives appear to have the fewest problems since neither forces the interviewees to focus on color. Avoiding priming respondents about color is important if the survey is to observe McCarthy’s guideline that the survey’s method must approximate market conditions as well as preserving the double-blind nature of the research. Consumers walking into a wine store are not asked to focus on color; they walk in and see a profusion of wine varieties, bottles, packages, brand names, and colors. If the VEUVE CLICQUOT ORANGE serves as a trademark, consumers will walk into the store, see the orange in the midst of the other cues and think, “VEUVE CLICQUOT.” This is what the survey aims to test. The second alternative, where the wine context is directly mentioned, also appears to have some advantages over the first for ease of administration so long as the broad category “wine” is used rather than “champagne.” For these reasons, it is likely to be adopted.

In any event, these considerations show the critical importance of the use of suitable placebo (control) colors (discussed below). Finally, this discussion highlights the way in which various design issues (recruitment, framing, interviewer supervision, stimuli and question design) are necessarily conflated, meaning that survey guidelines that deal with these issues, topic by topic, are insufficient to guide implementation.

C. Survey Method and Location

Fundamentally, the survey needs to show the color in some way and then test reactions to it. Showing people a photo or

57. See 6 McCarthy, supra note 1, § 32-163, at 32-335. See also supra note 19. See also Michael Rappeport, A Replication Problem in Survey Design, Including a Critique of the Decision in Thoip v. Disney, 100 TMR 1360 (2010). In the context of surveys designed to test for consumer confusion, he argues that consumers get information at the point of sale and in other contexts and the survey should test for the past, present, and future deployment of the trademark. But in contrast with these surveys, studies of trademark identification look to the past, not necessarily to the present or the future. The application for registration asserts that the mark already acts as a badge of origin. In this respect, such surveys should replicate as closely as possible the conditions in the time up to the application. Rivera makes a similar point. Rivera, supra note 12, at 688.
picture or swatch of color in an Internet survey\textsuperscript{58} has the problem that the specific color used may not be represented accurately. The VEUVE CLICQUOT product is also a three-dimensional object and, arguably, the survey should reflect this reality. This is again on the grounds that the survey must endeavour to approximate market conditions.\textsuperscript{59} People do not buy swatches of color, they buy physical products that have colored labels.

Given these considerations, a survey method that shows interviewees a physical object is required.\textsuperscript{60} The choice then comes down to three main alternatives: intercept interviews, a testing center, or in-home interviews.\textsuperscript{61} All are problematic.

With in-home interviews, a stratified random cluster sample based on contacting homes in person is possible, but given the low incidence of champagne buying and drinking, it would be especially expensive. In a random sample, the people who are eligible for interview are likely to be mobile and therefore hard to catch at home. They are also likely to have security systems designed to exclude people like survey researchers. The recruitment of the in-home interviewees is, therefore, more likely to come from a purchased list of wine buyers or champagne buyers. They would be contacted prior to interview; an appointment would be made and an interviewer would be sent. This in itself is an expensive process. Potential interviewees may well ask the source of the contact and, if it needs to be given, cueing as to category (see above) may be an issue. A list drawn from a general Internet panel or some other survey panel may be preferable. These panels often

\textsuperscript{58} Thornburg advocates the use of Internet surveys where possible on the grounds of cost, eliminating interviewer error, facilitating the use of controls and lessening the need for verification. The need to render color and show three-dimensional products, however, prevents the use of this technique for VEUVE CLICQUOT or indeed any product where a naturalistic setting for display is required. Thornburg \textit{supra} note 49, at 114-123.

\textsuperscript{59} Rivera, \textit{supra} note 12. Of course, to administer the survey there is always some element of artificiality. The testing cannot occur inside wine stores in sight of the champagne on display for fear that other cues are present. Similarly, consumers do not typically see wine bottles without brand names written on their labels. Within these constraints, as much realism as possible is indicated so that the survey results can usefully approximate market conditions.

\textsuperscript{60} Diamond observes that “in-person interviews generally are the preferred method of data collection, especially when visual materials must be shown . . . .” Diamond, \textit{supra} note 10, at 402.

\textsuperscript{61} In one case involving the secondary meaning of color, the Trademark Trial and Appeal Board gave some weight to a telephone survey in which interviewees were read a description of a purple sandpaper and asked whether it identified a particular company. See Saint-Gobain Corp. v. 3M Co., 90 U.S.P.Q.2d 1425 (T.T.A.B. 2007). However, this seems clearly inappropriate on two grounds. First to function as a color, a person needs to see it, particularly as the word “purple” can describe a range of colors and therefore it is unclear whether the mental image this question conjures up accurately represents the specific color for which registration is being sought. Second, this is entirely different from the real world situation in which a buyer sees a physical product.
collect information on a broad range of consumer behaviors,\textsuperscript{62} such as drinking or purchasing wine. While the source of the contact would be known, the specific product that was the basis for recruitment need not be revealed. Whether the Internet panel is representative of the consumer behavior of interest is another issue.\textsuperscript{63} Finally, a major problem with in-home interviews is the lack of effective field supervision.\textsuperscript{64}

Additionally, the dispersed nature of American cities means high costs for travelling time for in-home interviews. Of course, expense in itself is not a sufficient reason to avoid a particular method. There is no point in conducting a survey with a cheaper (but flawed) method. Arguably, the parties would be better off with no survey at all when the cost compromises the design. As will be shown below, a strict observance of the questionnaire format requires extensive interviewer training. Thus, apart from the cost, a real limitation of in-home surveying is that interviewers are difficult to supervise in such a situation, particularly as they have to administer a complex survey design. Finally, there is the practical difficulty of having to carry around a collection of bottles to show interviewees.

A testing center\textsuperscript{65} offers controlled conditions for bottle presentation, lighting, and display. Interviewees would be recruited from lists, as for in-home interviewing. However, the method requires interviewees to come to a location that may not be close, possibly restricting the representativeness of any sample. Assuming a relationship between wealth and champagne buying, the payment to attend may need to be very high to ensure an adequate sample is recruited. Costs per person would be even higher than recruiting people for a group discussion, but the sample size would be far larger, making this a highly expensive proposition. People could be invited to take part after using a short screening survey that disguises the product category and location of production. Those who reported, \textit{inter alia}, that they bought or consumed wine could be asked if they were willing to come to the interview center. This gives reasonable control over the priming effect. However, a testing center is a somewhat artificial environment, violating the requirement for a real world context.


\textsuperscript{64} Diamond \textit{supra} note 10, at 411-412. Validation (re-contacting a sample of interviewees) is one way of checking quality for in-home interviews but lacks the intense supervision possible in telephone interviewing or intercept interviews where supervisors are continuously present.

\textsuperscript{65} Also known as "central location surveys." See Thornburg, \textit{supra} note 49.
A third, cheaper alternative is to use intercept interviews in shopping precincts and near stand-alone liquor stores. These locations contain the type of artificial lighting likely to be used in liquor outlets, adding to the realism of what is shown. The intercept method itself may be critiqued on sampling grounds (discussed below), so a very careful design is needed to address this concern. In general, long questionnaires are at risk of producing poor data quality. However, the proposed survey format detailed below is quite short and suitable for an intercept location in which consumers are rushed for time and may be carrying heavy items from the store. Interviewers may be asked later to record their reactions to the survey, and interviewees may be asked about the survey process itself to address these concerns.

Because of the strict and complex requirements for sampling, survey stimulus presentations, and questionnaire design (see below), the survey may be critiqued on the grounds that its complexity may lead to errors in administration. Thus, appropriate training and supervision is critical. One approach to administering this complexity is to equip interviewers with handheld computers to administer the survey design and to capture responses. An independent auditor, separate from the company commissioned to conduct the fieldwork, may be engaged to observe and report on the survey's administration. The intercept method has been common in surveys accepted by the courts. This method has the advantage of relative ease of keeping a range of test packages nearby but out of sight. Close supervision is also easier with this method. However, the choice of malls, shopping precincts, and stand-alone liquor outlets is important, as discussed below.

D. The Stimuli Presented

Fundamentally, the survey is designed to test whether a relevant consumer, on seeing the VEUVE CLICQUOT ORANGE, is taken to the brand VEUVE CLICQUOT to the exclusion of other brands. That is, it tests whether the color has acquired a secondary meaning and acts as a trademark for VEUVE CLICQUOT. For trademarks like LACOSTE, where the mark is registered in black and white, this faithful rendition is not an issue. But with color, it is crucial. VEUVE CLICQUOT is registered not for all colors that a

---

66. Id. at 94-95. Thornburg notes that such approaches can offer a controlled environment but critiques the necessary quota sampling approaches that may be subject to self-selection. As noted below, there needs to be systematic screening of possible interviewees to select qualified people.

67. The research on the relationship between survey length and data quality is equivocal. See a brief review in Elisabeth Deutskens, Ko de Ruyter, Martin Weitzels & Paul Oosterveld, Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study, 15 Mkt. Letters 21 (2004).

68. Swann, supra note 3.
consumer might describe as orange, but for a particular color. In
VEUVE CLICQUOT's case, the color has no Pantone Matching
System (PMS) number, though PMS 137C is considered to be
close.69 The principles of stimulus generalization70 suggest that if a
particular color triggers a response from the consumer, then
similar but not identical colors may also trigger this response. This
may be relevant for trademark surveys concerned with passing-off
or deceptive conduct by competitors. But in terms of registering
colors as trademarks, the use of the precise color is important.
Colors rendered on three-dimensional objects and in shop
environments may look different from a flat swatch of color
observed in daylight. But the colors presented must be realistic in
terms of what consumers experience during the buying process;
hence, it is important to reproduce them accurately in the survey.

Trademark registration for color is more likely to be sought for
high-frequency purchases, specialist niche brands, or highly
promoted brands. Thus, one possible determinant of the response
is that interviewees are merely guessing a well-known or well-
publicized brand. This was one of the reasons why framing the
survey as being about "wine" is contentious. In order to assess this
tendency, a control (placebo) stimulus71 that is identical, except for
color, needs to be shown. If the color is to act as a trademark, the
VEUVE CLICQUOT ORANGE should produce a far higher brand
response than the response to the control color. Swann72 notes the
difficulties that arise in choosing an appropriate control stimulus,
but calls the process a "case-by-case conundrum." Diamond
suggests that a "control stimulus should not be less attractive than
the experimental stimulus if the survey is designed to measure
how familiar the experimental stimulus is to respondents."73 In the
proposed survey, the control color chosen must meet two
requirements that are possibly contradictory. First, the color must
not be in common use in the category or must at least be one that
is not strongly associated with any brand. Second, the color needs
to be plausible, that is, consumers would regard it as a possible
color used in the category. If the color were implausible, consumers
might react to the color and judge that no product in the category
would use it, hence weakening its ability to be a neutral control

69. David Young, Autumn Shades—the Colour Orange ... or is it Yellow?,
Marks Newsletter (Sept. 2006).

70. Werner K. Honig & Peter J. Urcuioli, The legacy of Guttman and Kalish (1956):
Twenty-five years of research on stimulus generalization, 36 J. Experimental Analysis

71. Swann, supra note 3.

72. Id. at 327.

73. Diamond, supra note 10, at 399.
stimulus. This phenomenon has been called “incoherence.” In the
case of a very well-known brand, it may be argued that a specific
category cue like “champagne” may trigger a familiar brand name,
which is then contradicted by the color being discrepant with the
brand name, that is, an incoherent combination. The VEUVE
Clicquot survey is, however, less likely to be subject to any
incoherence criticism, given the wide range of champagne brands
in a market where no brand is completely dominant and the fact
that only the broad “wine” category is used to recruit interviewees.
A pale blue solid color is plausible given the colors used to market
champagne, but not one that is associated with any particular
brand. An alternative such as green may also be suitable because
it is common on champagne brands. Some pre-testing of colors
associated with brands needs to be done if there is any doubt about
the control color being plausible, attractive, and coherent.

While the application for registration is likely to be based on a
single color, in reality, a variety of colors is likely to be used on
many products, some in combination with the main color. For
example, chocolate packages may use minor color to denote variety, such as green with peppermint chocolate.
While it is clear that one version shown to interviewees should use
the color in question on its own, as a baseline, other forms of
presentation with other colors may be required. For example, in
one jurisdiction, BP Petroleum sought registration for green alone
when it actually used other colors (white and gold) in its livery but
had its registration overturned on appeal partly because
the survey adduced did not test color combinations nor vary the
predominance of green. The orange in VEUVE Clicquot labels
is a solid color, with only some text in black and red written over
them. While the trademark application is for the color alone, the
color itself always appears in the context of other marks, signs, or
symbols. A relevant question, therefore, is whether the color
should be tested singly or in combination with these other indicia.
In the case of VEUVE Clicquot, labels with and

74. Ujwal Kayande, John H. Roberta, Gary L. Lilien & Duncan K.H. Fong, Mapping the
bounds of incoherence: How far can you go and how does it affect your brand?, 26 Mktg. Sci.
504 (2007).

of its commercial nature, market share data in the champagne market is very difficult to
obtain. This industry source was the best available. Where a party from the industry
commissioned the research, they would have access to trade data, which could be used to
provide for better estimates.

76. The survey evidence adduced in this case was criticized for not taking into account
that BP used green as a predominant rather than an exclusive color in its livery.
Woolworths Ltd v. BP Pte (No. 2), 154 FCR 97 (2006).

77. Thornburg discussed the need to isolate the effect of an individual indicium from
other signals of origin. The experimental design described in this article, by systematically
comparing the indicium alone with the indicium in combination with other stimuli, is able
to isolate this effect. Although Thornburg says there is a quandary because the indicium on
without text could be tested, should text be relevant. However, as 
mast, if not all, brand labels include the brand name, this is 
arguably unnecessary. In some product categories, color is not used 
everywhere, but as the dominant feature with other colors (e.g., 
white panels with the brand name on them). If this is a relevant 
issue, labels with and without solid blocks of color would need to be 
tested.

Although the trademark application is for the single color, 
stronger evidence may exist if packages that use the predominant 
color with minor use of other colors also produce the identification 
response. However, this may be over-elaborate for testing VEUVE 
CLICQUOT, as the label is orange throughout with very limited 
use of other colors. Arguably though, the solid uniform color across 
the VEUVE CLICQUOT label and box is also a cue as to brand, 
part from color. Again, the survey has to assess the response to 
the color, not its deployment. Thus, labels with and without large 
white sections on which branding or product information might be 
written should be included. VEUVE CLICQUOT uses orange as 
the main label on its bottles, on the neck, and on the boxes in 
which it is sometimes sold. Ideally, variations in the use of these 
would show the importance of orange on the different parts of the 
bottles by testing various combinations. However, this would add 
considerably to the expense for little gain. The main vehicle for 
consumer recognition is presumably the label or the box color so 
these variations may be unnecessary. Consideration should also be 
given to the shape and size of the label. Given that rectangular 
labels are common to the category, the survey could present 
participants with a rectangular label but in a different size from 
that used by VEUVE CLICQUOT. The different-sized label is used 
to avoid the criticism that the label size is itself giving a subtle 
cue as to identity.

Branded products do not simply contain a package with a 
single color used uniformly over it. In the case of champagne, there 
is usually a label attached to a dark glass bottle, normally in a 
traditional bottle shape. The shape itself may give some clues, 
particularly if VEUVE CLICQUOT has unique branding elements 
such as glass color or an embossed logo. If the impacts of these 
elements are to be tested in combination with color, various 
combinations of bottles will be needed, some of them originating 
with VEUVE CLICQUOT, others with non-VEUVE (that is, 
competitor) shapes. In this case, shape is not a key issue, so bottles 
from a competitor could be used. That will remove any suggestion 
that it is the shape/color interaction which is producing the 
identification response. In order to test generalizability to see
whether color serves as a badge of identity for a class of goods, different sizes of bottle might also be tested. Hence, two sizes, one a traditional bottle containing 700 mL, and another size, either a half-bottle or a magnum size, might be specified. In addition, champagne bottles usually include a neck label (a registered trademark for VEUVE CLICQUOT in some jurisdictions) and a colored foil wrapping over the neck of the bottle and wire-restrained cork. For the sake of the exercise, it is probably sufficient to have empty bottles without foil and a plain neck label not in any traditional VEUVE CLICQUOT shape. A champagne bottle shape, even if not VEUVE CLICQUOT, may be said to provide another form of framing. It would be possible to test the label on a different wine bottle shape. This raises the incoherence problem. VEUVE CLICQUOT packaged in a non-champagne bottle would presumably be an incoherent combination and, hence, suppress the VEUVE CLICQUOT response. There is some justification in also testing a wine box shape—a packaging device used by several types of wine, including champagne.

As a result of all these considerations, the choice of stimuli could consist of three package types (a box and two sizes of bottles), two colors (orange and light blue) and two label types (with and without a white section). Potentially, this would give twelve combinations, but they could be reduced by using only the labels with white sections on regular champagne bottles. This would reduce the number of packages to be used to eight, as Table 1 shows.

<table>
<thead>
<tr>
<th>Color - Orange</th>
<th>Package Combinations to Be Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orange</td>
<td>Solid label</td>
</tr>
<tr>
<td></td>
<td>700 mL bottle</td>
</tr>
<tr>
<td></td>
<td>375 mL bottle</td>
</tr>
<tr>
<td></td>
<td>box</td>
</tr>
<tr>
<td>Orange</td>
<td>Label with white section</td>
</tr>
<tr>
<td></td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Color - Blue</td>
<td>Solid label</td>
</tr>
<tr>
<td></td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>G</td>
</tr>
<tr>
<td>Color - Blue</td>
<td>Label with white section</td>
</tr>
<tr>
<td></td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

It could be argued that a complete factorial design, one using all twelve possible combinations, would be required. However, the proposed design allows for generalizability over package types and
sizes, control color versus the VEUVE CLICQUOT ORANGE, and solid versus mixed labels. Each interviewee should see only one stimulus, in order to avoid order effects. Within each gender/age quota in each center, the stimulus bottle shown should be randomly assigned to avoid any interaction effects. The stimuli should be numbered in a random manner, e.g., 997, 625, etc., so that the number is not correlated with any feature, meaning interviewers are given no cues about the significance of one package over the other.

E. Population and Sampling

A survey sample gives a type of evidence useful in legal cases that sworn statements from particular, selected consumers cannot. Surveys seek to represent the knowledge of the population of appropriate people. Ideally, the survey should use a random sampling design. In this approach, each person in the appropriate population has a known, finite chance of being selected. Because random sampling gives a method of estimating results in the population with a known degree of accuracy, this technique is preferred. Statistical tests, such as those which compare the identification response to VEUVE CLICQUOT ORANGE and the control color, assume random sampling. In practice though, non-random sampling methods, known as "non-probability sampling" methods, are commonly used in surveys adduced in legal proceedings, especially when "some tangible item needs to be shown to the respondent." As was argued above, a shopping precinct or mall that contains a champagne outlet or stand-alone liquor outlet is a highly appropriate location to find, recruit, and interview wine buyers. Under these circumstances, random sampling is not possible. Rather, a well-designed quota sample that specifies what groups of people are to be recruited, where and when, is required. So long as it is carefully constructed, with no evident biases, it should be acceptable to the courts.

In preparing this specific design, the first issue is to define the population. As noted, the most relevant are people who drink or buy wine. Given intercept interviews, the most likely place to find

---

78. In other words, all packages showing the VEUVE CLICQUOT ORANGE would need to show a similar response to provide evidence of distinctiveness. This is fundamentally the generalizability issue in a practical form.

79. Only in simple random samples is this probability of selection equal for all people. Strata samples may use different proportions for some groups. See Richard L. Scheaffer, William Mendenhall III & R. Lyman Ott, Elementary Survey Sampling (5th ed. 1996).


81. McCarthy describes the use of mall intercepts quota samples as being "sufficiently reliable to be admitted into evidence." 6 McCarthy, supra note 1, § 32:165, at 32-343.

82. Thornburg supra note 49, at 97.
these people is in shopping precincts that include wine outlets or stand-alone liquor stores. All these stores should be ones that sell champagne in large volumes or, if this cannot be determined, those that stock a range of champagnes. This has the effect of narrowing the sample frame to give a greater focus on places where champagne buyers shop. In defining the sample frame, the next step is to identify a range of diverse regions. They could be based on regions that have cities with high champagne sales, such as New York, San Francisco, and Miami.\textsuperscript{83} The two outlets in each location should be demographically matched so they could be directly compared, if necessary. In the United States, there are estimated to be at least 1097 liquor outlets\textsuperscript{84} for VEUVE CLICQUOT, apart from restaurants and bars. The choice of location in each city should be such that one outlet sells VEUVE CLICQUOT and a matching champagne outlet of similar type and location does not. The outlets are matched to avoid the criticism that the survey sampling is biased toward VEUVE CLICQUOT. The use of three major cities is designed to increase the heterogeneity of the sampling frame. Arguably, the survey should be extended to other areas to increase generalizability, but this would greatly increase costs. The fieldwork company conducting the survey would be given lists of suitable precincts, shopping malls, or outlets and replacements that could be made if some locations were unavailable for research. They would not be informed as to the rationale for the selection of outlets. Where interviews are conducted at stand-alone liquor outlets or in mall locations, the interviews must be conducted in a location out of sight of the champagnes on sale.

Already, this requires 6 outlets (3 regions by 2 types of outlet). Pragmatically, this may have to be cut down by eliminating one region like Miami. The more interviews are conducted per mall, the more cost-efficient is the data collection, but the more clustered the sample and, therefore, the lower the accuracy.\textsuperscript{85} Given this, there is a need to choose as many outlets as possible. The assumption is made in this article that at least 6 will be used. Even this may be criticized as being too few to represent the entire United States. If similar survey patterns are shown in the different locations, this criticism is easily deflected. If, however, large differences are found, this heterogeneity implies that a larger sample should have been used. It is, thus, important that the

\textsuperscript{83} Wine Institute, http://www.wineinstitute.org/resources/winefactsheets/article92. Arguably, the survey should go to more regions, but the cost would be extremely high.

\textsuperscript{84} Superpages, http://yellowpages.superpages.com. This was one of the few publically available sources for this information. If Veuve Clicquot or its legal advisers commissioned this survey, they would have access to their own internal list of distributors.

structure of the U.S. champagne market be well-understood and reflected in the choices of location. Ideally only those with higher levels of champagne sales would be chosen.

More outlets mean higher costs; for example, paying mall fees\(^{66}\) and briefing a wider group of interviewers. Given that some wine buyers or consumers would know little or nothing about champagne, this sample would need to be big enough to have a sub-group able to be analyzed separately, if relevant. In addition, given the need to show packages in various configurations (see stimulus choice below), these considerations apply to the sub-samples also.

Next, there is the issue of who to sample in each shopping location. Trademarks are to be registered for a class of goods, not a class of consumers. It is a legal issue as to which population group is relevant, but ultimately good sense would indicate that the minority of people who do not consume or buy alcohol\(^{87}\) should be excluded from the survey, even though some may be aware of high-status products like champagne. In the United States, the proportion of wine drinkers is lower than the number of beer drinkers,\(^{88}\) while the consumption of imported sparkling wine is less than 2% of the total wine consumption.\(^{89}\) Thus, even the group of alcohol buyers or consumers may be too wide, as some would have little knowledge of wines or champagnes. A further complication is that the use of the term “champagne” is limited to wines produced in the Champagne region of France.\(^{90}\) As VEUVE CLICQUOT comes from that region, it is appropriately called “champagne.” Many sparkling wines that use the “methode champenoise” of production could not legally be classified as champagne, even if consumers use this term to describe a wide variety of sparkling wines. Similarly, if the survey screening used a question to filter out those with little knowledge of champagne or sparkling wine, this would cue the interviewee as to the specific category. Even if interviewees are given a list of items they may have consumed or purchased, the interviewer knows which of the categories is relevant. Unintended communication at various points in the survey may be an issue. As noted above, the framing

\(^{66}\) Mall owners charge market research companies a fee for the use of their premises for research.


of the survey will affect how the sampling is organized. It is assumed that the survey is framed as being about wine. By focusing the survey on outlets that sell champagne, the survey becomes practical to implement, without the dangers of cueing a specific category.

Given the hypothetical 1996 application date, the most relevant group are those who legally bought or consumed wine in 1996 or before,\textsuperscript{91} making them 37 or older in 2012. The explicit statement of 37 being the minimum age for recruitment is unlikely to cue either interviewers and interviewees about the time of filing issue. However, if this were considered a danger, the minimum age could be set at 33 and the people aged 33 to 36 screened out at the analysis stage. Assuming this is unnecessary, the population of interest could be defined as people 37 and over who buy or consume wine. Quota groups 37 to 48, 49 to 59, and 60 and over should be specified. The older quota group is important because it could be argued that older people frequent shopping outlets less often. Gender, climate, race, income, and occupation are also likely to affect the buying and use of wine.\textsuperscript{92} While climate can be accounted for by choosing varying locations (as above), it would not be possible to have quota controls for all the other demographic variables.

One likely approach is to set location quotas by gender and age. Within each combination of gender and age, the number of people would be in approximate proportion to the population of wine drinkers and purchasers. Table 2 shows how the quotas could be organized. Over the whole sample, there could be minimum quotas for White, Hispanic, Black, and other Americans. Finally, there may be a pattern of times when particular groups attend shopping malls (e.g., on the way home after work). Interviews should, therefore, be spread through the interviewing period to increase heterogeneity. Otherwise, the survey may be criticized on the ground that it is biased toward frequent shoppers.

Taking this into account, with at least 6 gender-age combinations and eight types of packages to be shown (see above), there are 48 quota combinations in each location (see Table 2). It is assumed that they are evenly split between the age groups and genders, but this could be modified if industry data suggested other proportions better reflected underlying purchase and consumption patterns. The number in each quota for each mall

\textsuperscript{91} This could include current and lapsed buyers of champagne. Potential users may not be as relevant, given the 1996 filing date.

should be chosen on the basis of the minimum effect size anticipated for the trademark element. For example, suppose that at least a 15 percent difference between the experimental and control groups is needed to argue for the effect of VEUVE CLICQUOT orange when compared with the control color. Just what evidence a court would accept as important is open to argument (see below), but any effect smaller than that would be trivial. Using a test for differences in proportions, it can be calculated that a sample size of at least 96 for each package type tested would be needed. For example, at least 96 people should be shown orange boxes and 96 people the blue boxes. With 6 age-gender quotas and 8 package types, there should be a minimum of 6 people in each quota group. Thus, the overall sample size to be recruited would be 1,728 (6 locations by 6 quotas by 6 people by 8 package types) at a minimum. Such an approach lowers the risk of a biased sample that is too dependent on any single type of location.\textsuperscript{93} Although locations are proposed which are likely to include many people who buy or consume champagne, it may actually be prudent to have a sample size greater than this.

Response rates are a relevant issue in all surveys. Low rates may compromise the representativeness of the resulting sample.\textsuperscript{94} A proper way of defining the response rate is to look at the ratio of obtained interviews to eligible interviewees who have been contacted.\textsuperscript{95} This is hard to determine in this case. First, there is the issue of whether eligible interviewees include all wine purchasers or just those who purchase champagne. Second, it is only after screening that it could be determined whether a person was eligible and by this stage they would have already consented. The practical solution is to keep records of those who were contacted and did not cooperate and those who cooperated but were screened out. In this way, some crude indication of response rate can be estimated. These difficulties reinforce the importance of widely recruiting people in different locations, in different quota groups, and at different times of the day or week.

\textsuperscript{93} Diamond, \textit{supra} note 10, at 356.

\textsuperscript{94} Id. at 383-386.

\textsuperscript{95} Mike Shaw, David Bednall & John Hall, \textit{A Proposal for a Comprehensive Response-Rate Measure (CRRM) for Survey Research}, 18 J. Mktg. Mgmt. 533, 538-540 (2002). Diamond observes the need to collect and disclose this information in the survey report. Diamond \textit{supra}, note 10, at 415.
Table 2
Per Mall Quotas

<table>
<thead>
<tr>
<th>Package Type</th>
<th>Males</th>
<th>Females</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>37-48</td>
<td>49-59</td>
<td>60 and over</td>
</tr>
<tr>
<td>A</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>H</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>48</strong></td>
<td><strong>48</strong></td>
<td><strong>48</strong></td>
</tr>
</tbody>
</table>

One criticism of this sampling approach is that intercept interviews are inherently clustered, reducing the effective sample size. Another criticism is that the selection of outlets may not be representative of the whole population of champagne buyers. Additionally, people who frequent shopping precincts may not fully represent the population of interest. For example, buyers of rival brands (who may be contesting registration) may claim that their buyers who use other channels to buy champagne—for example, via the Internet or direct ordering or in clubs and bars—are not represented. A final objection may be that the selection is biased toward frequent patrons of the liquor outlets. The survey researcher may point to the considerable care taken to establish a heterogeneous sample broadly representative of the population of interest. Ultimately, though, no practical sample is ever likely to be perfect. Post-weighting might help to fine-tune the survey estimates if there are known characteristics of sparkling wine purchasers and drinkers.\(^{96}\)

People in each gender-age group would need to be randomly allocated in each location to one of the four experimental or four control groups. Randomization is necessary to avoid any systematic bias in assigning packages to quota samples and is a standard feature of any experimental design.\(^{97}\) It ensures comparability between the various control and experimental groups.

---

96. While it could be argued that the figures should be weighted back to the general population, this does not define the target market of those people who buy or drink champagne.

97. See Cook & Campbell, supra note 47.
F. Questions to Be Asked

1. Use of Open-Ended Questions

Pre-coded questions are problematic because they signal to the interviewer what answers are expected. There is, therefore, a risk that the interviewer may then inadvertently signal to the interviewee what answer is expected. This is exactly contrary to the double-blind nature of the research described above. Belson and Duncan long ago showed that checklists with specific answers suppress responses to open-ended "other (specify)" options, compared with open-ended questions without checklists with specific answers. 98 Similarly, respondents also tend to choose items at the top of the pre-coded list. 99 Whether interviewers are subject to the same issues is unknown, but there is a risk that they may inadvertently induce interviewees to suppress responses to questions with checklist answers or to choose items at the top of a list. In any event, using pre-codes with open-ended questions puts further performance pressures on interviewers who have to code answers while dealing with all the other tasks in the survey. Research has shown that this in-field coding can lead to inferior data quality. 100

Other than simple "yes/no" type questions, the pre-coding of responses further removes courts from the actual words used by interviewees, compounding the concern about hearsay evidence. Thus, Swann 101 asserts that, wherever possible, open-ended questions should be used to avoid leading questions that suggest responses either to interviewers or interviewees. McCarthy has observed that open-ended questions are not without their own problems, as questions asked about brands may tend to favor well-known or dominant brands. 102 This is why one or more control groups are required. However, he also observed that open-ended


100. Edith de Leeuw, Joop J. Hox & Don A. Dillman, International Handbook of Survey Methodology (European Association of Methodology) 213 (Taylor & Francis, 2008). They observe that the interviewer has to determine in which response category the answer can be placed which is a less reliable method to a doubly coded response back in the fieldwork company. Additionally, the original answers of the respondent are subsequently no longer available for independent analysis or for production to the court. Similar concerns are raised by Robert M. Groves, Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer & Roger Tourangeau, Survey Methodology 335 (2d ed. 2009).

101. Swann, supra note 3.

102. § McCarthy, supra note 1, § 32:172, at 32-369.
questions provide the courts with insights unavailable through numeric survey results. "The respondents' verbatim responses to "why" questions may provide a window into consumer thought processes in a way that mere statistical data cannot."103

Interviewers need to be instructed to record the answers to the questions completely and accurately; in other words, verbatim responses need to be recorded. Interviewers should be explicitly warned not to probe responses to open-ended questions. Probing involves asking further questions in order to understand or expand the answers given. Because probes are often used in survey research, interviewers need explicit instructions and training not to use them. Their use could be seen as continuing questioning until a sought-after or expected response is obtained.

2. Question Formats for Trademark Surveys

While Swann details two standard formats for trademark questions, the "Eveready" and "Squirt" tests,104 neither is suitable, as both deal with possible confusion, not with the badge of origin question. Similarly, an empirical study of trademark distinctiveness by Lee and his colleagues asked consumers whether a word mark served as a source indicator.105 Although this has the merit of seeing whether the trademark element takes the consumer to the brand, framing the survey as being about color has its own inherent disadvantages, so the technique was not applicable. A similar formulation in secondary meaning cases was suggested by Jones.106 He proposed that the indicia (such as a "mark, product, or package feature") be "... printed on a display board ... removing all other indicia of source from the packaging" and then asking a question such as, "Do you associate [brand name product] with the [generic product name] of one, or more than one, company?" Or, "Does [brand name product] connote one or more than one source of [generic product name]?" Just how these questions would incorporate color is unclear. One possible formulation is, "Do you associate this color with the champagne of one, or more than one, company?" Or, "Does this color connote one or more than one source of champagne?" Jones notes that a follow-up question "which company?" should be used. Thornburg107 suggested a similar wording: "Do you associate the claimed

103. 6 McCarthy, supra note 1, § 32:178, at 32-387.
104. Swann, supra note 3.
trademark with a product/good of one or more than one company?” This assumes awareness of the claim and knowledge of what constitutes a trademark. This latter point is especially contentious.

Palladino later proposed a similar approach to surveys testing secondary meaning using the formulation, “If this type of product were made by more than one company, would you associate [the word at issue] with the product of one, or more than one, company?” In the current context, this could translate as follows, “Champagne is made by more than one company. Would you associate this color with the champagne of one, or more than one, company?” Both the Jones and Palladino formats are simple and go from the indicium to the source. However, there are several issues. The first is the framing of the domain as “Champagne.” It has already been noted that the framing of the survey topic and questions affects both the recruitment of interviewees and the form of the questions that should be asked. Second, they both focus on color in the question and in that sense what is proposed runs counter to a realistic depiction of the buying situation in which a consumer goes into a store and sees champagne in a variety of get-ups, but on seeing the VEUVE CLICQUOT ORANGE, identifies the brand.

A possible alternative is to show a champagne bottle with a label in the appropriate color and ask, “Do you associate this wine with one, or more than one, company?” However, the consumer may not know the anonymous source (i.e., the company) behind the brand. While VEUVE CLICQUOT (the brand) and Veuve Clicquot Ponsardin (the company) are closely connected in this case, it cannot be assumed that consumers will necessarily have knowledge of the company behind the brand. The question, therefore, must allow for the possibility that the consumer knows the brand, the company, or the product and is not necessarily sure which is which.

The word “associate” is another key issue. Just because a particular indicium is strongly associated with a brand does not mean that it serves to identify the source. A brown-color fluid is commonly associated with COCA-COLA, but this element, although part of every COCA-COLA drink and therefore strongly associated with it, cannot serve as a trademark. Similarly,


111. 1 McCarthy, supra note 1, § 7:48, at 7-137.
survey evidence that showed that when consumers were given the
name of a telephone directory they could name the color of the
directory was not accepted as providing evidence of secondary
meaning. "The fact that the consumers surveyed knew that their
SPRINT yellow page directories were red does not require the
conclusion that red telephone directory covers had become a source
identifier for those consumers." 112 It is acknowledged that in most
jurisdictions, "associate" has been the test of distinctiveness.
McCarthy113 observes that "It is the word 'association' which
appears most often in judicial definitions of secondary meaning by
both federal and state courts."

This may mark a point where there is a disjunction between
the words the courts use to describe the secondary meaning
concept and the words that are best suited in a survey to test it.
Association is a symmetric concept. WHISKAS cat food is
associated with purple, and purple is associated with WHISKAS.114
However, the purple may identify WHISKAS, but WHISKAS does
not identify purple. The identification relationship is not
symmetrical. It goes from the indicium, purple, to the source. This
precisely matches the secondary meaning concept. Hence a test of
secondary meaning as measured by a question asking about
"association" may be insufficient to evidence the relationship
between the symbol and its source.

It is possible to conceive a situation where consumers may
only be able to nominate a single source in relation to a color and
the association question, but would not give the same answer to a
question about identification. For example, consumers may say
they associate a dark purple color with only one chocolate brand,
CADBURY. But when asked the identification question, "Does this
color identify the products of one or more than one company, or
not?" they may say "no" or "don't know" because they are unsure
whether other brands use this color and therefore could not be
certain that it identified CADBURY exclusively. As Palladino
observes, "By identifying the goods of one producer and
distinguishing them from the goods of others, a trademark
indicates the source of goods."115 It is therefore argued that the
word "identify" is the best representation of the secondary
meaning concept. This directly tests the mental pathway from the
color to the brand. All this assumes that consumers would
understand the word "identify" in the sense intended and would, in
principle, be able to distinguish these responses from those to the

112. R.H. Donnelley Inc. v. USA Northland Directories, Inc., 2004 WL 2713248 (D.
113. 2 McCarthy supra note 1, § 15:5, at 15-13.
114. Golder & Montalto, supra note 33.
115. Palladino, supra note 108, at 858.
word "associate," "Identify" is clearly a more common word than "connote," which Jones suggested. As with all questionnaires, some form of pre-testing is necessary\textsuperscript{116} to check that all contentious questions are understood and that problems with administration are identified. However, as will be argued below, normal pilot testing is insufficient to test for the validity of the questions used in surveys conducted for legal (or indeed other) purposes.

3. Basic Question Sequences

It is proposed that upon showing the stimulus that the interviewee be asked immediately if he or she can identify the source. This matches a real life shopping situation in which a consumer confronts a display of products and spontaneously identifies the brand. "Here is some wine. Does this wine identify any particular brands, products, or companies to you, or not?"\textsuperscript{117} This question has the virtue of not forcing the interviewee to focus on any one aspect of the package and especially not color. It also avoids implying that only a single "brand, product, or company" should be named or that a single response is expected. Any form of question that suggests there is only a single source is leading in the context of a survey designed to test for distinctiveness. The "brands, products, or companies" formulation is designed to allow for the likely possibility that consumers will not necessarily know the "anonymous source"\textsuperscript{118} of the indicium—that is, they may be unaware of the company that owns or produces the item. The question format invites a simple "yes/no" response. The use of the "or not" part of the question is clumsy but is designed to avoid the suggestion that a "yes" response is expected.\textsuperscript{119} Some interviewees may spontaneously offer brand, product, or company names in answer to this question. Nonetheless, a follow-up question, "Which brands, products, or companies?" would then be asked of all interviewees. This question is open-ended, and no pre-codes would be provided to the interviewers, nor would probes be allowed.\textsuperscript{120}

\textsuperscript{116} William Foddy, Constructing Questions for Interviews and Questionnaires 185-186 (1993).

\textsuperscript{117} A less directive opening question such as, "Please look at this wine. Can you tell me what, if anything, comes to mind?" is possible but would in itself produce problems and therefore is not proposed. If the subsequent identification question is asked following the opening general question, interviewees may feel they have already answered the question, artificially lowering their identification response. Again, there is also the danger that interviewees will be directed down semantic pathways that are irrelevant to the survey's purpose. In addition, the survey analysis (see below) needs to find some way of combining the identification responses to both questions.

\textsuperscript{118} 2 McCarthy, supra note 1, § 15-8, at 15-20 to 15-23.

\textsuperscript{119} Diamond notes the dangers of an acquiescence bias even with this simple form of questioning. Diamond, supra note 10, at 394.

\textsuperscript{120} Gough, supra note 12. Diamond notes the difficulties with probes being applied consistently and the need to record both the response and the additional interviewer
However, open-ended responses will need to be coded, as discussed below. While other questioning approaches may also be possible, it is argued that the critical element in a trademark survey of this kind is to show the indicium of interest and see whether it results in the person identifying the source. This approach embodies that principle.

After the opening two questions, there is, arguably, a need for further questions to ask about the reasons for the identification responses. Strictly speaking, if there is a strong difference between the VEUVE CLICQUOT ORANGE and its control color in the identification question, that may be enough to establish the distinctiveness of the color. However, further information about the reasons for the responses given is likely to be useful to the tribunal or court in deciding whether to register the color. Thus, a follow-up question appears relevant. As it is possible that more than one “brand, product, or company” has been identified, a separate loop of subsequent questions for each one mentioned needs to be devised. The questionnaire or computer program (if a handheld computer is used) needs to allow for a potentially unlimited number of specific responses. This complicates the administration of the questionnaire, meaning that good training and supervision are vital. The interviewer must effectively make a judgment about what specific responses have been given. Answers such as “that Veuve company,” “the champagne company in Reims,” “Edouard Brun,” “Moët & Chandon,” “Marguet Père et Fils,” or “Gloria Ferrer” (a local brand) need to be recorded separately. The follow-up question would take the following form for each one identified, “Why do you say this identifies (name of brand, product, or company)?” Again, this is an open-ended question, with no probing, with the answers to be recorded verbatim. Exhibit 1 shows the basic question sequence that would be used.

As noted when discussing interviewees’ understanding of the word “identify,” some form of pre-testing is essential. Normally, test interviews are conducted that check the administration of the

questioning used. It is argued, though, that even this is not sufficient, since interviewers need to be instructed when to stop or what response is sufficient. Even if applied consistently, which is itself difficult, the creating of expectations as to what is a relevant or sufficient answer has the potential to bias the process. Diamond, supra note 10, at 394-395.

121. Diamond, supra note 10, at 395. Diamond notes the possible issues raised by order effects. The survey design is at pains to ensure that neither the recruiting nor framing of the survey has an effect on answers to the key survey questions. Beyond that, the survey merely shows a stimulus and then asks the key questions. One of this article’s authors, in a trademark survey involving a colored chocolate pack, sequentially showed two packages to interviewees, not one. This was designed to increase the cost effectiveness of the survey by collecting more data. The showing of the packages was systematically rotated to minimize overall order effects. However, the survey results did show order effects, suggesting that once the person had been exposed to the survey task, this affected his or her responses to the subsequent stimulus. This reinforces the need for a short, simple set of questions.
questionnaire and allow the interviewers to observe whether people are having difficulty answering. It is argued that this is insufficient. Rather, pre-testing using cognitive interviewing principles122 would help establish the validity of this, by showing that interviewees interpreted it correctly. In essence, this approach allows the questionnaire to be administered as it would in a regular survey. The interviewer then conducts an in-depth interview with each interviewee, going back over each question. Interviewees are asked to explain in their own words what each question means to them and are then asked to describe the thinking process that led to the answer they gave. A verbatim recording of responses here may itself provide useful supplementary evidence. Where a crucial word such as “identify” is contentious, cognitive interviewing would provide the court or tribunal a more substantial evidentiary base from which to evaluate the survey.

4. Asking Questions About Time

The next set of questions can be especially contentious. At the time of application, the applicant was asserting that the mark already served as a badge of origin. Hence the most relevant survey evidence related to consumer beliefs at the time of application, not the time at which the survey was conducted. Ideally, the VEUVE CLICQUOT survey should have been conducted at around the time registration was being sought. For a whole host of practical reasons, the research may take place some years after the event. This is especially likely if an earlier decision about registration is appealed by either the company or its opponents. Thus, after the event, the court or tribunal will still need to make a judgement on the issue. So, the survey needs a sufficiently reliable way of estimating consumer knowledge at the time registration was sought.

---

122. Gordon B. Willis, Cognitive Interviewing: A Tool for Improving Questionnaire Design (2005). Such testing, going under various different names, has a long history in social research going back at least as far as Belson’s seminal work, William A. Belson, The Design and Understanding of Survey Questions (1981) and others. While it may be asserted that interviewees do not have complete insight into how they arrived at their responses, the technique yields useful information about how words like “identification” are understood. Additionally, useful insight may be gained into how questions are interpreted, the conscious reasoning behind the answers or what the interviewee thought was the expected answer. At the same time, the method itself is prone to errors in questioning and interpretation, see Frederick G. Conrad & Johnny Blair, Sources of Error in Cognitive Interviewing, 73 Pub. Op. Q. 32 (2009). Tourangeau et al. point to a lack of standardization in the use of cognitive interviewing and the need for coding schemes to address systematically the comprehension, memory load, judgment, and response issues common in survey research. Roger Tourangeau, Lance J. Rips & Kenneth Rasinski, The Psychology of Survey Response 326-328 (2000). Thus like all social research methods used in a legal domain, cognitive interviewing needs to be carefully designed and applied.
Consumer estimates involving the passage of time are often unreliable, especially if the event lacks salience or involvement. One simple method is to ask, “You say that [name of features] identify/identifies [name of brand, product or company] to you? Can you say for how long [name of features] have/has identified [name of brand, product, or company] to you?” followed by a question, “Why do you say that?” Again, no pre-coding should be used because of accusations of expectation and bias if date ranges that coincide with periods before and after the registration application are supplied. Administratively, this further sequence needs to be added to the loop of questions following each specific response.

The approach suggested may yield some useful information, but there are many issues to be confronted. For marketing purposes, companies often relaunch or refurbish brands. Even if the color in question has been used for a long time, its deployment may have changed. Thus, if the brand has not consistently used its trademark element in the time since registration, some interference with brand knowledge may be expected and time estimates made more difficult. In the case of VEUVE CLICQUOT, this is not an issue, as there has been continuity over a long period. Second, the question has to be asked in a way that allows interviewees to say they “don’t know,” rather than feeling obliged to give an answer. Assuming that interviewees can make reasonably confident responses, they are likely to be prone to a systematic bias known as “telescoping.” Originally, telescoping was seen as always involving a systematic underestimation of the passage of time. This accords with the everyday experience that time has often passed more quickly than first realized. This is known as “forward telescoping” and is well-documented. Researchers have also identified backward telescoping, where the interviewee overestimates the passage of time. While such research outcomes are observed less often, some studies have found them to be equally likely.


124. See also Foddy, supra note 116, at 93, and Zaichowski, supra note 110, at 54. Zaichowski argues that although particular colors are strongly associated with particular brands, this is a low involvement issue for consumers.

125. Woolworths Ltd v. BP Plc (No. 2), 154 FCR 97 (2006). At the time of the survey, BP was beginning to change its livery by using thinner strips of green, more white strips, and a new logo.


127. Auriat, supra note 123.
Memory research has uncovered some techniques to improve the accuracy of recall; these techniques have potential to improve the accuracy and insights available from the simple questions outlined above. The first of these techniques is to establish a landmark event and then relate it to the estimate of interest. For example, most people can remember the 9/11 tragedy and specifically what they were doing on that day. Similarly, a form of time banding might also be used to constrain estimates, for example, before, after, or during the time Bill Clinton was President. The problem with these methods is the way consumers develop brand identification. One sighting of an orange champagne bottle may not be enough for consumers to identify it as VEUVE CLICQUOT. Repeated exposures may be necessary for the consumer to come to use orange as the badge of identity. Precisely when consumers come to regard orange as a badge of origin is, thus, difficult to pin down. The problem with this approach then is that there is nothing in the consumers’ semantic networks that would link the landmark or time-banded event to acquiring knowledge of VEUVE CLICQUOT.

In the case of VEUVE CLICQUOT, another useful approach for measuring recall would be to tap into episodic memories. For example, a bride’s family may choose VEUVE CLICQUOT for her wedding reception, making the event and date memorable. “You mentioned [Name of company, product, or brand]. Can you recall any particular occasion where [Name of company, product, or brand] was served?” If the answer is “yes,” then “When was that event?” should be asked. Of course, not all consumers will have memorable events involving VEUVE CLICQUOT, and, for those who do, the memorable event may have occurred after the relevant filing date. Possibly, the questionnaire could record the year when significant events like a wedding or graduation occurred so that the identification response can be tied to those dates. All these significant events would need to be tested, as specifying a relevant date may signal to the interviewee what is expected. These methods should, ideally, be empirically pre-tested to see what, if any, reliable evidence is likely to be produced. Whether the additional effort would yield enough consumers with these precise memories is very questionable.

It may be that all that can be done is to get broad estimates of time. If the estimates stretch considerably back beyond the filing date, despite their imprecision, evidence has been provided that consumers have held their judgment for a long time and probably many of them held it at the filing date. “If the results of the survey need to be translated back to an earlier date, consider additional

forms of evidence that support that proposition." In the case of VEUVE CLICQUOT, the survey evidence will provide, at best, a useful complement to other evidence: "... the most compelling evidence is likely to be found elsewhere, in the archived sales and marketing data, the clips of old TV ads, the old posters — ... which show that the activities of the business regarding the mark before the relevant date, the market and the consumers were much the same as they have been since then. Such traditional evidence is essential to support a submission that although the passage of time augmented the reputation of the mark by ... [the survey date] and some judicious discounting of the strong survey results is appropriate, there is every reason to believe that in substance the reputation was "similar and sufficient" at the relevant date."

5. Overall Question Design Issues

The specific design in terms of questions to be asked, their sequence, and their coding have revealed many problematic issues. In order to suggest a possible question sequence, Exhibit 1 shows a minimum set of questions. Pre-testing of any proposed sequence of questions is needed to ensure that the questions are understood in a way that matches the legal requirements for relevance and reliability. Pragmatically, pre-testing may also give the party planning to commission a full survey some guidance about whether to proceed.

G. Interviewers

The surveys are to be administered by interviewers. There are several practical issues here. First, the interviewers need to be well-trained and well-briefed. Surveys for trademark purposes are becoming more and more complex in design—for example, they may include a partial factorial design, as proposed in this article. It is common in commercial market research for interviewers to be given some latitude in tailoring their approach to suit potential interviewees in both the recruitment preamble and in the conversation between questions. Such tendencies need to be controlled. Similarly, interviewees need to be trained to record answers accurately and not to interpret answers. Thus, high levels of training are necessary. Ideally, briefings should be recorded. Care also has to be taken that neither the interviewers nor the

129. Gough, supra note 12, at 415. See also 6 McCarthy, supra note 1, § 32:171, at 32:361 to 32:362. McCarthy noted that a survey about a past period should be not ignored but “given whatever weight the decision-maker thinks is appropriate.”

130. Gough & Downes, supra note 34.

131. Willis, supra note 122.

field supervisors are briefed on the ultimate client or purpose of the survey. For this reason, it is desirable that the survey expert rather than solicitors for Veuve Clicquot commission the fieldwork.

**Exhibit 1**

**Minimum Question Sequence**

| Q1. Here is some wine. Does this wine identify any particular brands, products, or companies to you, or not? |
| Yes | 1 ASK Q2 |
| No  | 2 GO TO DEMOGRAPHICS |
| Don't know | 3 GO TO DEMOGRAPHICS |

Q2. IF YES TO Q2, ASK, Which brands, products, or companies? RECORD VERBATIM. DO NOT PROMPT. DO NOT PROBE.

Q3. FOR EACH BRAND, PRODUCT, OR COMPANY, ASK, Why do you say this wine identifies [NAME OF BRAND, PRODUCT, OR COMPANY]? RECORD VERBATIM. DO NOT PROBE.

Q4. You say [NAME OF FEATURE(S) FROM Q3] identify/identifies [NAME OF BRAND, PRODUCT OR COMPANY] to you. Can you say for how long [NAME OF FEATURE(S) FROM Q3] have/has identified [NAME OF BRAND, PRODUCT, OR COMPANY] to you? RECORD VERBATIM. DO NOT PROBE.

Q5. Why do you say that? RECORD VERBATIM. DO NOT PROBE.

In this hypothetical VEUVE CLICQUOT study, one of the benefits of intercept interviewing will be the continual presence of a field supervisor as part of the interviewing team. Ideally, some form of independent monitoring is also needed to certify that interviewing occurred at the places and times specified, that the various stimulus objects were properly controlled, and that the questionnaire was properly administered away from any VEUVE CLICQUOT or other wine outlet. Video recording of interviews would be ideal, but it would add considerably to the time and cost and could affect the nature of the interview. Privacy concerns would also be harder to address. To ease the burdens on interviewers who are administering a complex series of conditions, quotas, and questions, handheld computers can be used to program the sequence of question loops and to record answers. This requires appropriate training. Use of handheld computers has
a further advantage in that data can be reported back as collected, assisting the fieldwork company with quota controls on demographics such as race that are applied to the survey overall rather than to specific interviewing locations. Thornburg has observed that mall intercept surveys, along with other approaches that use interviewers to collect data, “are all subject to being discredited and devalued due to the risk that survey interviewers could falsify or mischaracterize data entries, or information could be given without sufficient verification.” To minimize these risks, the strict controls, supervision, and independent testing outlined above are necessary. While Thornburg promoted the use of Internet surveys, it was previously argued that these are unsuitable in the case of trademark surveys related to color.

**H. Coding**

The use of open-ended questions normally requires the answers to be coded in order to summarize the results. If the coders know the purpose of the survey, they may be inclined to highlight or otherwise bias the way the coding categories are developed and deployed. One solution is to get a coder to devise a code frame without receiving instructions as to its structure or the purpose of the survey, in order to preserve the double-blind nature of the survey process. While this will be considered unusual by a fieldwork company, the nature of the answers usually makes this a relatively straightforward task. Then, two coders using the code frame independently of each other would help ensure the reliability of the coding task. Where any inconsistencies are identified, the coders meet to resolve the differences, carefully recording cases where they occur.

An issue with coding is that it puts a further hearsay step in the path from interviewee to the court or tribunal. That is, it involves someone else reporting on what other persons have said. Thus, it is necessary to have available for inspection the verbatim answers to questions. This allows any party to devise their own code frame or to test the robustness of the frame used to produce the survey results in the expert’s report.

**I. Analysis and Reporting**

In terms of analysis, simple inferential statistics are probably sufficient. A crucial test is the extent to which the VEUVE CLICQUOT ORANGE leads to the “VEUVE CLICQUOT” identification response. While an earlier review suggested that at

---

133. Thornburg, *supra* note 49, at 91. However, Thornburg observes that, properly conducted, this type of survey is “still generally a reliable and effective method of trademark surveying.” *Id.* at 95.
least 25 percent must "associate the mark with a single source,"134 others have claimed that a 75 percent or 80 percent "association" is needed to establish secondary meaning of a trademark.135 After reviewing many past cases, McCarthy asserts that, "Generally, figures over 50% are regarded as clearly sufficient."136 Similarly, Palladino asserted that "over half" nominating that a product came from a single company would provide sufficient evidence of secondary meaning.137

A second test is that the identification should be considerably higher than the comparable response for the control (placebo) color.138 While it is arguable as to what this level of difference should be, a distinction needs to be made between statistical significance and a real-world difference. Suppose that, across all four orange packages, of the 300 people who drink or buy champagne, 83 percent say that the orange package identifies VEUVE CLICQUOT, compared with 76 percent for the 300 champagne drinkers who saw the control pack. The difference is statistically significant, but in real-world terms, 76 percent is a very high placebo response, suggesting that something other than color is largely determining both sets of responses. A third test may be expressed as a high signal-to-noise ratio. That is, the ratio of people identifying the VEUVE CLICQUOT brand in question should be far higher than for any other specific brand when "orange" is shown.

The survey analysis has to allow for multiple responses, for example, multiple brands named in response to the identification question. Modern survey software can easily cope with multiple responses. However, it does mean that the expert is sometimes reporting on responses and sometimes basing the analysis on people, which can be disconcerting to readers not used to this type of survey evidence.139

The expert’s report on the survey needs to be sufficiently detailed so that no findings are hidden yet it must be clear enough so that legal practitioners not versed in survey methodology and statistical analysis can follow it. Any perceived attempt to report

134. 6 McCarthy, supra note 1, § 32:100, at 32-413. Jones, supra note 106, at 476.
136. 6 McCarthy, supra note 1, § 32:100, at 32-414.
137.  Palladino, supra note 108, at 870. Here, Palladino is citing his 1994 TMR article (Surveying Secondary Meaning, 84 TMR 155 (1994)).
138.  See Saint-Gobain Corp. v. 3M Co., supra note 61. Here it was argued that a 7 percent difference between the responses to the color and the control was insufficient to evidence distinctiveness.
139.  It has been argued that survey results should always be reported in terms of the percentage of interviewees giving any particular answer. However, sometimes the focus is on answers given to questions. Where multiple responses are possible, great care is needed to make clear the nature of the results.
results in a way that favors the party who sponsored the survey is likely to be condemned. Unlike a survey conducted for public policy or commercial clients, minimalist discussion and only a few, key implications are included. It is tempting for the expert who has designed the study not to acknowledge any weaknesses in the design, its implementation, or reporting; however, if registration is opposed, the opposing parties will find these flaws. Outright denial or clever but unreasonable argument will reduce the expert’s credibility and limit the usefulness to the court or tribunal of the information collected. On the other hand, experts employed by opponents to critique the survey may themselves criticize the survey in unreasonable ways. For example, they may claim that a quota sample is insufficient to represent the population, that framing the topic or questions is biased, or that more package combinations should have been tested. Although better samples and more variations would add value, they are impractical in terms of cost, producing only incremental gains in relevant information.

Given likely opposition to the survey findings, it would make sense for the party that is commissioning the survey to engage two independent experts to critique the VEUVE CLICQUOT survey. The first expert should be engaged at the design stage, in order to question the design and to suggest alternatives. The second expert should be engaged to review the draft report. It is perhaps unwise to have one person fulfill both roles. Once the outside expert has influenced the design, they are likely to be more comfortable with the analysis and reporting than someone seeing the project for the first time.

If the litigation reaches court, the survey expert will face rigorous cross-examination. In at least one jurisdiction, opposing experts may actually appear together in an effort to resolve differences or at least to identify issues on which there is fundamental disagreement, a process described as “hot-tubbing.”

V. CONCLUSIONS

Using survey research in trademark proceedings involving secondary meaning is now common, and the survey designs are becoming increasingly sophisticated. General principles about how to design and implement this research provide a useful framework but are not specific enough to guide the development of particular surveys. As this article has shown, various conflicting issues need to be resolved if such surveys are to be sufficiently robust to withstand intense scrutiny by parties seeking to discredit findings they find contrary to their interests. This article has laid out in

detail how such a survey might be conducted by describing key implementation issues relevant to both the legal and research worlds. Although the focus has been on the VEUVE CLICQUOT brand and its distinguishing color, ORANGE, the issues identified have a general application to all secondary meaning trademark surveys. Thus, the article has shown how the “devil in the detail” of implementation has a direct impact on the defensibility of the evidence produced by the survey. The resulting question design is actually simple, but the overall survey is clearly complex and expensive in terms of recruiting people and in testing various package alternatives.

This article has also highlighted how the various elements of the design necessarily interact. Simple guidelines that take these topics in isolation are clearly insufficient. Rather, an analysis suggests that a decision logic is needed for surveys of secondary meaning:

1. Is the trademark indicium in color? If the answer is “Yes,” either singly or in combination with other indicia like logos or text, then the indicium needs to be shown in person to the interviewee in a realistic situation. Indirect methods such as using the Internet are not possible because there is no control regarding how the color is rendered on-screen. The next issue is how the color is deployed. For example, if a service company mostly uses color on flat surfaces, such as in written communications to customers, walls in retail outlets or on brochures, then survey stimuli using pictures are suitable. Color photographs can then be used, assuming that the color is accurately rendered. This may open the possibility of mail surveys, or to surveys with interviewers carrying around photos. Such interviews are theoretically possible in homes, as well as in intercept interviews. However, where colors are on actual physical product packages, mocked up packages rather than photographs are more realistic. Colors under these circumstances vary in shading, depending on their angle to the light source. Lighting should match the circumstances in which buyers typically encounter the packages. This suggests that lighting typically encountered in shopping malls or specialty stores where the products are sold is a necessary environment for conducting the research.

If the answer is “No,” then rendering of the trademark indicia on a computer screen may be possible, opening the doorway to far cheaper, quicker, and more accessible Internet surveys. This assumes that flat surfaces (logos on clothing, for example, or in the screen or paper communications of service providers) are a realistic representation of the product class in use. Internet surveys
using rented panels are not necessarily representative of the population unless they impose rigorous quota sample designs. Quota samples, such as the one identified in this article, specify the groups of interest to be surveyed. These quota groups should be broadly representative of the population. In addition, quotas should be filled progressively through the sample period, so that the heterogeneity of the population is represented. For example, if Internet survey interviews are conducted during the first morning they are posted online, they will not represent people who get home later from work. Similarly, quota samples in shopping precincts need to spread out interviews over the week to capture diversity. Some obscure groups may be hard to find. However, they do offer great control for the experimental manipulation of the stimulus materials. In addition, there are no interviewer effects to worry about and no training required. Again, if the rendering of the trademark indicia or their normal deployment is on three-dimensional surfaces, personal interviews will be required in an appropriate lighting setting.

2. How should the survey topic be framed at the recruitment phase and at the start of the interview? Essentially the framing of the survey topic is problematic if it is too wide or too narrow. It is therefore proposed that the survey be framed in terms of a category that includes, but is not limited to, the product or service class in question. Wine as a super category that includes champagne would be acceptable, as would telecommunications for broadband services. This corresponds to a normal purchasing environment where the consumer is likely to encounter the product or service among others of a similar class.

3. How should the sample frame be defined? The population of interest certainly includes the people who purchase or may purchase the product or service and perhaps those who use it as well. It may also include the people who influence that purchase, as purchasers may be buying on behalf of others. For example, a daughter may choose VEUVE CLICQUOT for her wedding but may ask her father to arrange the purchase. As the recommender groups are harder to identify, it may be simpler to define the core population of interest as being buyers. If these groups do not treat the indicium as a trademark, then it is unlikely that the reactions of any other group would identify the trademark in sufficient numbers to convince a court.

Within the buyer population, a frame for sampling must be defined. In home surveys, which are still possible in many
countries, it would be advisable to use a form of multistage cluster samples. These will be possible only where the product or service is in high use and access to homes is feasible. With the Internet, quota samples are needed. Similarly, intercept interviews in shopping precincts require quota samples, based on a frame of outlets that fairly represents the population of interest. This article gives a detailed example of how this might be implemented. Finally, random assignment of people within quota categories to the various experimental and control groups is a necessary condition for the experiment.

4. How many variants of the experimental and control stimuli are required? The trademark indicium will typically be deployed in a range of packages and labels for products, or in a variety of surfaces and communications if a service is involved. If the findings of the survey are to be of use to the court, then they need to be capable of generalization to a product class, not just a single product. A decision needs to be made as to how many variations on package types, packaging, and labelling are required. With color, in particular, it is rare that the color covers the entire surface of a package or label. Again, in order to be useful to the court, the color may need to be tested on its own and in other configurations. For example, chocolate may make minor use of other colors to indicate variety—red if it includes cherries, brown if it has a caramel filling, and so on. Combinations of these with and without other package features, such as ingredient labels, may need to be tested. Testing of indicia such as logos may be simpler. But, again, the experiment should take account of the variety of situations in which these logos are deployed, in order to generalize the findings.

The control stimuli need to be plausible, but not strongly associated with any particular brand. As shown, this is tricky for color. It should be simpler for logos or stylized text. As with the indicia and deployment types for the test stimuli, so too should there be matching control stimuli. It would appear unnecessary to use a full factorial design, but a broad enough range of combinations should be tested so that several direct comparisons can be made.

5. How is the identification response to be measured? The crucial test is whether the indicium takes the consumer to

141. Scheaffer et al., supra note 79, at 335-360. Traditionally, these are based on a sample of census districts, the selection of a cluster of homes within these, then the selection of a person to interview from each household. Call-backs are normally required to interview people who are out or unavailable for interview at the time of contact.
the brand. The key decision is whether the word used in the question, such as “identify” actually is understood in the way intended or whether some other form of words is required to provide evidence that courts would accept. This is, in part, an empirical question that requires pretesting both with the experimental and control stimuli. Next, follow-up questions that allow people to give reasons for their response may be necessary. The issue of certainty or near certainty in these responses has not been examined and may need testing. If consumers commonly give a brand response like VEUVE CLICQUOT but express a degree of uncertainty, then there must be doubt about the identification response. Again, pre-testing is essential if such a question is to be asked.

6. How are responses to be recorded? As noted, open-ended questions without probing are required. This means that interviews need to write down the responses verbatim. If the answers are straightforward, presumably this is easy to record accurately. Again, some testing of this, possibly by recording interviews, would help. Internet surveys record what interviewees type in—but it is uncertain that interviewees take the trouble to express their full thoughts.

7. What survey controls are in place? The designer needs to build in checks for interviewers to confirm their instructions, to check that they stood in the right place when recruiting people, and to make certain they used the survey questions in precisely the manner specified. Similarly, the coding of open-ended questions needs to be checked.

There are also decisions to be made as to what level of identification, both absolute and relative to the control, is an acceptable level of evidence that an indicium acts as a trademark. Essentially, this is a legal issue, but the discussion in this article suggests that at least a majority of the target market should make this identification and that this should considerably exceed the response to the control color.

Finally, the article has discussed the issue of time estimates relevant to the time the registration application was lodged. The issues here are extremely problematic. Unless there is some highly memorable event or time band associated with the brand in question, only approximate time estimates are likely. These may still be of assistance to the court or tribunal when combined with other evidence of the use of the indicium by the brand.

Given all these problems and criticisms, parties may wonder why they should expose themselves to the risk and cost of a consumer survey. Arguably they would only do so if they were reasonably certain that a properly conducted, reliable, and
unbiased survey would produce helpful information. With a product like VEUVE CLICQUOT, directed at a premium niche market, the difficulty and expense in finding suitable people to research would be especially problematic. Presumably this is why Veuve Clicquot Ponsardin relied on more traditional forms of evidence to support its registration applications in Europe and elsewhere. Should the company wish to register their color in the United States, a lack of systematic survey evidence may count against it if it is feasible to conduct a survey and if the cost of doing so is affordable.\textsuperscript{142} However, if they ever do need a consumer survey, they will at least have some specific design guidance from this article. More generally, the article identifies specific but complex design decision issues that will need to be addressed in any trademark survey focused on secondary meaning.

\textsuperscript{142} Evans & Gunn, supra note 8, at 4, 25.