EXPERIMENTAL DESIGN AND THE SELECTION OF CONTROLS IN TRADEMARK AND DECEPTIVE ADVERTISING SURVEYS*

By Jacob Jacoby, Ph.D. **

I. INTRODUCTION

Whether commissioned on behalf of plaintiff or defendant, consumer surveys have become a common feature of trademark and deceptive advertising litigation, with the number of surveys actually proffered into evidence representing the tip of the iceberg. One reason is that some surveys conducted in anticipation of being offered into evidence fail to yield the desired results. Other surveys, although shared with one's adversary, never reach court because they become a factor in achieving settlement. The basic point: surveys are increasingly being relied upon as both a litigation and settlement tool.

Spurred by their increasing prevalence and also, in part, by the Daubert trilogy,[1] intellectual property counsel who commission surveys and the courts who evaluate them are becoming more sophisticated in regard to surveys. One way in which this has been manifested is by an emerging understanding that "controls"[2] often are called for in litigation surveys. However, despite such general recognition, many counsel and courts have only the barest understanding of the scientific logic underlying controls, what

---

* Many intellectual property attorneys become involved in both trademark and advertising litigation. The title of this article reflects the fact that, with few exceptions, the basic principles regarding experimental design and the selection of controls apply equally to surveys in both domains. Portions of this article have been drawn from the following copyrighted sources: Jacob Jacoby, Experimental Designs in Deceptive Advertising and Claim Substantiation Research, in Advances in Claims Substantiation 119-41 (Cynthia M. Hampton-Sosa ed., 1991); Jacob Jacoby (in preparation), The Role of The Judiciary In Creating And Fostering Junk Science: A Scientist's Perspective.

** President, Jacob Jacoby Research, Inc., Associate Member of the International Trademark Association; member of the Editorial Board of The Trademark Reporter®. Jacob Jacoby holds an endowed chair as Merchants Council Professor of Consumer Behavior and Retail Management, Leonard N. Stern School of Business, New York University where, among other subjects, he teaches research methodology to Ph.D. students in the Departments of Marketing, Management and Organizational Behavior, and Computer Information Systems. He is also a Fellow of NYU's Center for Law and Business.

1. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993); General Electric Co. v. Joiner, 118 S. Ct. 512 (1997); Kumho Tire Co., Ltd. et al. v. Carmichael et al., 526 U.S. 137 (1999).

2. Although they take many forms, it is helpful to think of a "control" as a comparison used to rule out plausible alternative explanations for the observed effect.

controls are or may be, when their use is and is not called for, the considerations involved in developing controls, whether alternatives to controls exist, and other related considerations. These issues provide the focus for this article.

II. WHENCE THE FOCUS ON CAUSAL ISSUES?

Many concepts central to trademark and deceptive advertising law refer to psychological states held by the relevant public. Such concepts include secondary meaning, acquired distinctiveness, confusion, genericness, fame, blurring, tarnishment and deception.[3] With regard to some of these concepts, the pertinent question of law has nothing whatever to do with causation but, more basically, with the presence (and, if present, condition) of such mental states. One needs to establish that a mark, trade name, trade dress, slogan, etc. seeking protection *has acquired* secondary meaning, or *is* "famous," not what *caused* it to be so.

In other instances, however, the pertinent question is essentially one of causation. This focus on causation stems directly from the language of the Lanham Act. According to Section 43(a)(1)(A), action can be taken against an entity that uses something in commerce that "is likely to *cause* confusion, or to *cause* mistake, or to deceive. . . ."[4] Similarly according Section 43(c)(1), action can be taken when an entity "*causes* dilution of the distinctive quality of a mark. . . ." For example, in the typical infringement matter, simply demonstrating that a substantial proportion of the relevant public is confused by another's use of a mark, trade dress, etc., is generally a necessary but not sufficient condition for courts to grant injunctive relief. For confusion to be considered actionable, it must be shown to be *caused by* the allegedly infringing action(s), not by something else.

When scientific research is used to assess a causal proposition, and when a gatekeeper or counsel needs to evaluate the adequacy of such research, it becomes important to ask: How do scientists—particularly social scientists (inasmuch as the trademark concepts at issue refer to psychological states of mind)—go about assessing

3. For a more extensive discussion, see Jacob Jacoby, The Psychological Foundations of Trademark Law: Secondary Meaning, Genericism, Fame, Confusion and Dilution. 91 TMR 1013 (2001). Also see related discussion in Jerre B. Swann, Sr., Dilution Redefined for the Year 2002, 92 TMR 585 (2002).

4. Emphasis added. Similar language and emphasis is to be found in the Federal Trade Commission's definition of deceptive advertising and in the definition of misleading prescription drug advertising developed for the Food and Drug Administration in 1974. See Jacob Jacoby, Defining and Measuring Misleading Advertising (1974), Final report to the Division of Drug Advertising, Food and Drug Administration. A briefer description is provided in Jacob Jacoby and Constance B. Small, The FDA Approach to Defining Misleading Advertising, 39 Journal of Marketing 65-68 (1975).

causation? Actually, scientists employ several different strategies for assessing causation.

Although the scientist's approach tends to be more refined and rigorous, one strategy shared in common with lay people is observation. If three people testify they saw Mr. X deliver a single hard punch to Mr. Y's nose and Mr. Y, previously without blemish, now had a twisted nose protruding through the skin with blood spewing from it, it is likely we would not need a doctor's report to conclude that the broken nose and bleeding were caused by the punch to the nose. Observation would be sufficient. We recognize that no scientific controls are needed to arrive at this conclusion. For reasons discussed in greater detail below, some types of causal phenomena are so basic and obvious that controls should not be necessary.

Another considerably more sophisticated scientific approach that does not necessarily rely on controls is known as Structural Equation Modeling (SEM). Relying on correlational (that is, non-experimental, non-control)[5] designs, "SEM frameworks encompass data analytic methods that can be used to analyze any data and to test any hypothesis. In some ways, most of the most popularly applied statistical methods are simply special cases of SEM. SEM is powerful because in many situations it can explicitly take into account measurement error."[6] However, insofar as I know, SEM designs have never been used in assessing causal questions such as: Did the trade dress of Package A cause Person P to be confused?

The most common form of causal assessment strategy used across the sciences relies on experimental designs. Although the word "experimental" may conjure up impressions of things tentative or in an early stage of development (e.g., experimental aircraft), as used across the sciences, the term "experimental design" refers to well-developed and rigorous data gathering and

---

5. Understanding the distinction between correlational and experimental designs requires a foundation in statistics. Despite calls by, among others, Judge Richard Posner, for attorneys to receive such training as part of their law school education, such foundation remains lacking (see Richard A. Posner, The Problems of Jurisprudence, 468 (1990)) and providing such is well beyond the scope of this article. However, as a crude surrogate for such background, consider the following. We all understand that just because a rooster crows and, a few minutes later, the sun comes up, it does not mean that the rooster *caused* the sun to come up, even if this happens 100 days in a row. All this means is that there is an exceptionally high correlation between the rooster crowing and the sun coming up. The general rule is: Demonstrating a high correlation does not equate to demonstrating causation. On the other hand, relying upon the analysis of an interwoven set of correlations, SEM designs enable one to tease out causal implications.

6. Personal communication from James J. Jaccard, Distinguished Professor of Psychology, State University of New York at Albany. Prof. Jaccard is the author of several statistics texts. For treatments of structural equation modeling, see R. Hoyle, Structural Equation Modeling: Concepts, Issues and Applications (1995); K. Kelloway, Using LISREL for Structural Equation Modeling: A Researcher's Guide (1998).

data analysis procedures. The latter, which come into play after the former have been applied, refer to a family of statistical techniques[7] bearing names such as "one-way analysis-of-variance," "factorial designs," "randomized block designs" "analysis-of-covariance," "Latin Square designs" and "Graeco-Latin Square designs." Insofar as my experience is concerned, such statistical designs for analyzing data have yet to be introduced or commented upon in intellectual property litigation. What has received considerable attention, however, are experimental designs as data gathering strategies.

Experimental designs qua data gathering strategies is a subject touched upon, among other places, in several chapters of the Federal Judicial Center's (FJC) Reference Manual on Scientific Evidence.[8] Because this subject quickly becomes complex (with entire books written on the subject and doctoral programs often offering full semester courses on the subject), of necessity, and as the authors of these chapters would themselves readily acknowledge, the few pages in the FJC's Reference Manual are only able to supply a simplistic, incomplete, scratch-the-surface treatment. Although still at a relatively simple level, this article provides a fuller treatment of some fundamental issues in experimental design and how these fundamentals apply to intellectual property surveys. Discussion has been simplified to be appropriate for the intelligent non-scientist. As the subject is not easy reading for most non-scientists upon first exposure, the import of these concepts becomes apparent only after re-reading and serious study. However, an improved understanding of these issues will enable the reader to better differentiate between junk science and quality science.

III. A PRIMER ON EXPERIMENTAL DESIGN

Though experimental designs vary considerably in sophistication, all involve two fundamental components, namely: (1) the introduction of an event and (2) subsequent assessment of the presumed impact of that event on one or more other factors. Do we want to know if increasing air pressure in a particular balloon will *cause* the balloon to burst? Why don't we continue blowing up the balloon and find out? (Note, once again, a situation where a

7. Although the classic in this area is R. A. Fisher, The Design of Experiments (1935), since then, dozens of statistics texts have included the phrase "experimental design" in their title.

8. See David H. Kaye and David A. Freedman, Reference Guide on Statistics, in Reference Manual on Scientific Evidence 83, 90-96 (Federal Judicial Center ed., 2d ed 2000) [hereinafter Kaye & Freedman, Reference Guide on Statistics]; Shari Seidman Diamond, Reference Guide on Survey Research, in Reference Manual on Scientific Evidence 229, 256-60 (Federal Judicial Center ed., 2d ed. 2000) [hereinafter Diamond, Reference Guide on Survey Research].

control is not needed.) Do we want to know if an ad will *cause* people to be more interested in buying the product being advertised? Why don't we show some people the ad and find out? Do we want to know if a particular trade dress is likely to *cause* marketplace confusion? Why don't we expose relevant consumers to the trade dress and find out. The planning of an experiment begins with considerations such as these. As experimental designs can become quite complex, a special vocabulary and notation system has evolved to simplify description and discussion. These need to be explained before proceeding.

A. The Vocabulary and Notation of Experimentation

Experiments are designed to determine whether a *presumed cause* will generate a predicted or *observed effect.* The presumed cause is referred to either as the stimulus, treatment, test or independent variable. Those exposed to the presumed cause are called "subjects" (or "respondents" when experimental design is incorporated in a survey) and are said to comprise the *experimental*, *treatment* or *test group.* The effect is called the response, outcome, or dependent variable because its appearance is presumed to be activated by and dependent upon the prior appearance of the presumed cause. By convention, interventions of the treatment (or presumed cause) are denoted by the letter X, while the resultant effect is denoted by the letter Y.[9]

The basic notation of experimentation includes, at a minimum, two additional letters, O and R. By custom, the letter O (which stands for Observation) refers to the act of measuring the dependent variable. The letter R requires further explanation.

A fundamental feature of experiments is that they involve comparisons of various sorts, for example, comparing what happens after exposure to the presumed cause with what happens after non-exposure to the presumed cause (sometimes combined with exposure to something else). Such alternatives to the treatment are termed "controls." As described below, the purpose of a control is to enable us to rule out one or more alternative explanations for the observed effect so that the presumed cause remains as the only (or, more generally, the most reasonable) explanation for the obtained effect.

Within (Internal) vs. Between (External) Controls. Comparisons can be created either by assigning different people to the Experimental and Control groups, or by exposing the same person

9. In anything other than simple experiments, there may be more than one treatment condition or group. Under these circumstances, the various Xs usually have either numerical or alphabetical subscripts to differentiate one treatment from another. For example, X_D might be used to designate a treatment group exposed to a typical "disclaimer" while X_{MD} might refer to a group exposed to a "modified disclaimer" treatment.

to both Test and Control conditions. Suppose we used 200 subjects to examine the effect of blood-alcohol level on the ability to perform eye-hand coordination activities. We could assign 100 subjects to the Experimental group and have them imbibe a sufficient quantity of alcoholic beverages to raise their blood-alcohol levels to 10 percent. The other 100 subjects assigned to the Control group either would imbibe nothing, or would imbibe the same number of ounces of a non-alcoholic beverage (e.g., orange juice). After waiting an appropriate amount of time (say, 30 minutes), respondents in both groups would be tested on eye-hand coordination activities, and the scores of the two groups would be compared. If, on average, the control group performed significantly better, we might conclude that a blood-alcohol level of 10 percent causes deterioration in eye-hand coordination. Alternatively, each of the 200 subjects could be used as his or her own control—that is, each could be tested under both types of conditions (of course, having enough time elapse between the two experiences so that there would be no residual effects attributable to having consumed alcoholic beverages). In the former two-group case, we have used "External" controls (or, in statistical parlance, a Between-group design). In the latter single group case, we have used "Internal" controls (or, in statistical parlance, a Within-group design).

Common forms of internal controls used in trademark and deceptive advertising studies include other questions (such as those asked about meanings presumably not contained in an allegedly misleading ad) and other stimuli, as is often the case where respondents are asked about an array of items (products, ads, etc.). Some experiments involve the presence of both Internal and External controls. Examples of the use of internal, external and internal + external controls in trademark and deceptive advertising studies are provided in a later section.

Random Assignment. In the case of a "between-group" design, the subjects could be assigned to the Experimental and Control groups either randomly or non-randomly. Similarly, in the case of a "within-group" design, when each subject is given the Experimental (alcohol) experience and when each is given the Control (non-alcohol) experience could be determined either randomly or non-randomly. By convention, the letter R is used to signify when random assignment takes place.[10] Random assignment does not guarantee that the Experimental and Control groups will, on average, be equivalent in terms of their pre-exposure levels on the dependent variable of interest. Average equivalence across randomly formed Experimental and Control

10. "Random assignment" should not be confused with "random selection," the latter being the conceptual foundation for probability (as opposed to non-probability) sampling designs. See Jacob Jacoby and Amy H. Handlin, Non-probability Sampling Designs for Litigation Surveys, 81 TMR 169 (1991).

groups is an assumption (widely accepted across the scientific community) that, in the overwhelming majority of instances, turns out to be correct.

Some authors use the letter E to represent an Experimental group and the letter C to represent a Control group. Last, while participants in surveys generally are called "respondents," participants in experiments generally are called "subjects." When surveys utilize experimental designs, either term may be used.

B. The Logic of Experimentation: Ruling Out Rival Explanations

The principal function of experimental designs is to enable one to rule out alternative explanations for the observed effect so that we can conclude with a reasonable degree of scientific certainty that X caused Y, or that X caused Y for the reasons we say it did. Consider the question that typically provides the foundation for a Section 43(a) deceptive advertising action: Does exposure to a particular advertisement (or advertising claim) cause deception? Suppose after exposing someone to the ad, we observe (measure) that they are deceived. Does this mean that exposure to the ad *caused* this deception? It might; then again, it might not. The fact that the presumed cause is related to the observed effect (Y) does not necessarily mean that the two are causally related. (It can be conclusively shown that children with bigger feet spell better. Should we, therefore, stretch children's feet in an effort to "cause" better spelling? No, because the relationship is correlational, not causal. Children with bigger feet spell better because they are older.) Suppose this deceptive belief was present before we made our assessment, but we did not know this because we failed to check. Or suppose some other factor, not the ad, caused this deception (our spouse had misinformed us about the product), and exposure to the ad happened to be coincidental. Or suppose the measuring instrument was faulty so that, although it indicated deception, no such deception was actually present. Or suppose exposure to the ad actually caused an impact on some third factor (Z), and it was this other factor that "caused" the deception. For example, suppose exposure to the ad caused increased discussions about the product with others and it was misinformation obtained from these conversations that actually caused the deception. Such possibilities are called "rival explanations."[11]

Rival explanations are always available to account for why a study might suggest that a presumed cause and observed effect were causally related when, in point of fact, they were not. The greater the number of plausible rival explanations that can be ruled out, the greater the confidence we can have in inferring that

11. Plausible rival explanations are also called "threats to validity."

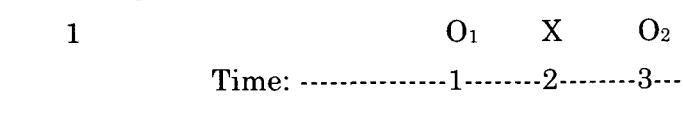
the observed relationship is indeed causal for the reason(s) we say it is. Generally speaking, experimental designs represent the best available scientific strategy for ruling out plausible alternative explanations.

As a basis for explaining how experimental designs are devised, let us return to the question: Does exposure to a particular advertisement (or advertising claim) cause deception? Suppose that instead of testing a single individual, the investigator selected a group of individuals to be representative of the population of interest, had them view the advertisement, then measured their level of deception. Since these events occur over time, they may be depicted along a time line, as in Figure 1, unfolding from left to right. Above point 1 is the symbol X, representing exposure to the advertisement at that point in time. Above point 2 is the symbol O representing the subsequent observation or measurement of deception. Often termed the "one group, post-test only" design, Figure 1 depicts the most basic design. Let us consider its efficacy.

Figure 1: Design 1

Suppose we found that, after exposure to the ad, many subjects were deceived. As discussed, this would not necessarily mean that the advertisement caused this deception, as the level of deception may have been just as high *prior* to exposure to the advertisement. Somehow, this rival explanation has to be ruled out. Design 2, as depicted in Figure 2, seeks to do just that. Here, as indicated by the O_1 and O_2 designations, deception is measured both before and after exposure to the advertisement. Doing so enables us to determine whether (and, if so, just which and how many) subjects came to the assessment situation already deceived. While Design 1 provides no basis for comparison, Design 2 enables us to compare "pre-test" with "post-test" observations. Although not an independent Control group, this pre- versus post-comparison enables us to gain some insight into the impact of the advertisement.

Figure 2: Design 2



Unfortunately, the situation is complicated by confounding factors[12] that render any conclusion of causation suspect. Even if we observe no deception at Time 1 and deception at Time 3, we could not necessarily conclude it was exposure to the advertisement (X) that caused this observed effect. Another potential explanation is that deception might have increased even in the absence of exposure to the ad. If the post-test observation was made one week after the pre-test observation, it is possible for something to have occurred during that week to cause deception, independent of the advertisement. To rule out this rival explanation, the researcher either could administer the test immediately after exposure to the ad or use a separate Control group—a group of respondents who had not been exposed to the contested advertisement and whose level of deception was also measured at Times 1 and 3. Using the absence of an X to indicate that the Control group was not exposed to the ad, this is depicted as Design 3 (see Figure 3).

Figure 3: Design 3

Group			
Experimental (Group 1)	O_1	X	O_2
Control (Group 2)	O_1		O_2

Time: ----------1--------2--------3-------- →

Using Design 3, the investigator would compare the level (if any) of deception for the Control group (Group 2) to the level of deception for the Experimental group (Group 1). If there were little or no difference between the two, it would be difficult to conclude that the advertisement was responsible for causing the level of deception observed in the Experimental group. In contrast, if the level of deception in the Experimental group was appreciably higher than the level found with the Control group, this would be consistent with an interpretation that the advertisement caused deception.

Unfortunately, even with Design 3, other rival explanations remain. Subjects in the Experimental group had their level of deception measured prior to being exposed to the ad. Perhaps sensitized by this measurement, the subjects approached the allegedly deceptive ad with a different mind-set then they normally would have, so that they were "primed" to be affected by the deceptive nature of the ad. It was this "priming," when coupled with exposure to the ad, not the ad itself, that caused subjects in the Experimental group to be affected by the deceptive elements,

12. A "confound" is another factor that creates an interpretive ambiguity. See Kaye & Freedman, Reference Guide on Statistics, supra n.8 at Section II.C.2.

and this effect is what was captured by the observation made at Time 3. It could be argued that had there been no such priming, exposure to the ad might have exerted no deceptive effect. Despite being similarly primed by the observation made at Time 1, because the Control group was not exposed to the ad, there would be no reason for them to exhibit deception on the post-test (Time 3). Hence, while mindlessly comparing the findings from the Experimental group at Time 3 with those of the Control group at Time 3 would suggest the ad had caused deception, this easily might not have been the case. Design 3 has no way of parsing out this rival explanation—a form of "measurement reactivity" termed a "testing-by-treatment interaction"—from the explanation that it was the advertisement, by itself, that caused the deception.

One solution to the testing-by-treatment interaction is to eliminate the pretest for both the Experimental and Control groups. This yields Design 4 (see Figure 4). By comparing the level of deception of subjects exposed to the advertisement with that of subjects who have not been exposed to the advertisement, the investigator might gain insights into the effectiveness of the advertisement. Unfortunately, Design 4 cannot rule out yet another rival explanation: Perhaps the differences in deception detected between the two groups was not caused by the advertisement, but rather was due to the fact that the two groups differed in deception to begin with. If one tries to accommodate such criticism by incorporating pretests, then we come back to Design 3 and the problem of the testing-by-treatment interaction.

Figure 4: Design 4

Group		
Experimental	X	O
Control		O

Time: ----------1--------2--------- →

The most sensible way to handle the above rival explanation is to *randomly* assign subjects to either the Experimental or Control group. Such a procedure would be unlikely to produce a situation where all people who were alike in one way (e.g., all held deceptive beliefs prior to participating in the study) were placed into one group, while all people who were alike in another way (all held no deceptive beliefs) were placed into the other group. Because it involves random assignment, although it may superficially appear to be the same as Design 4, from a scientific perspective, this design is substantially different and may be depicted as Design 5 (see Figure 5). By randomly assigning subjects to the two groups, it is unlikely that the *average* level of deceptiveness would be

different in the two groups. Thus, comparing the average post-exposure deception measure taken with the Experimental subjects to the average for the Control subjects would provide insight into the effect of the advertisement, without being confounded by the problem of testing effects or the problem of testing-by-treatment interactions.

Figure 5: Design 5

Group	Assignment		
Experimental	R	X	O
Control	R		O
	Time: ----------1----------2--------- →		

Some may find Design 5 to be counterintuitive because it assesses *change* on a dependent variable (level of deception) by measuring that dependent variable only *once* for each individual. Our natural inclination might be to assess the dependent variable first, expose the individual to the treatment, then measure the dependent variable again to see if and how it changed. As the above discussion indicates, there are problems with this strategy. Design 5 is an improvement in that it circumvents many of these problems. Yet it does have weaknesses. Because it uses a single assessment per individual, Design 5 only permits us to make statements about the effects of a treatment on average, for *groups* of individuals. For example, applying Design 5 to the case of evaluating the advertisement, we would be able to make statements such as "on average, deception was higher for the subjects exposed to the advertisement." In this case, we do not have the ability to make statements about whether, or by how much, the advertisement changed the level of deception for any given individual. Doing so would require both pre-test and post-test measures on the same individual.

If one wanted to obtain individual estimates of change and also "check" the effectiveness of random assignment, one might use Design 6 (essentially, Design 3 with the added feature of random assignment; see Figure 6). Unfortunately, Design 6 is a two-edged sword. On the one hand, it enables us to rule out the possibility that the presumed effect would have existed prior to the presumed cause. On the other hand, the pre-test assessment (at Time 1) may be "reactive," that is, it may sensitize the test respondents so that they experience the treatment (at Time 2) differently than they normally would. As a consequence, observing an effect (at Time 3) might be due to the interaction between experiencing the pre-test (at Time 1) and being exposed to the treatment (at Time 2). Under such circumstances, we have no ability to parse out whether it was

the pre-test, exposure to the treatment or a combination of the two that was responsible for the observed effect (at Time 3). This problem is exemplified by the Federal Trade Commission's study in FTC v. Kraft.[13]

Figure 6: Design 6

Group	Assignment			
Experimental	R	O_1	X	O_2
Control	R	O_1		O_2
	Time: ----------1----------2----------3------- →			

To this point, the designs outlined are quite simple, referring only to a single presumed cause (which, if present, exists in a single variation and at a single level of intensity) and the use of a single Control group. However, these designs are the building blocks for scores of more sophisticated experimental designs used to assess either a single presumed cause,[14] or to accommodate greater complexity in the presumed cause. While discussion of these factors is beyond our present scope, sufficient background has been provided to enable discussion of other pertinent issues. Perhaps none is more fundamental than the distinction between Fully Experimental vs. Quasi-Experimental Designs.

---

13. See infra Part V.E.

14. One particularly noteworthy example is the Solomon Four Group Design devised by combining Designs 5 and 6 and depicted here as Design 7. As the name implies, this design consists of four groups, two Experimental and two Control, to which subjects are randomly assigned. All subjects in one Experimental and one Control group are observed (measured) at Time 1; subjects in the two other groups are not measured at that time. At Time 2, all subjects in both Experimental groups are exposed to the treatment (the deceptive ad), while none of the subjects assigned to either Control group is exposed to the treatment. At Time 3, the respondents in all four groups are observed (measured). While this more sophisticated design makes it possible to rule out a greater number of potential alternative explanations than do Designs 5 or 6 by themselves, implementation of the Solomon Four Groups approach quickly leads to a proliferation of groups as the number of potential causal factors that need to be examined increases.

Design 7

Group	Assignment			
Experimental	R		X	O_1
Control	R			O_1
Experimental	R	O_1	X	O_2
Control	R	O_1		O_2
	Time:----------1----------2----------3-------- →			

C. Fully Experimental vs. Quasi-Experimental Designs

To qualify as a "fully experimental" design, an experiment must possess three key features: (1) the presentation or insertion of the independent variable (the presumed cause) must be under the experimenter's control; (2) there must be a comparison between what occurs to the dependant variable after presentation of the presumed cause versus what occurs to the dependant variable when the presumed cause has not been presented, and (3) there must be random assignment of (a) respondents to groups (in the case of between-group designs), or (b) occasions on which the presumed cause is present versus absent (in the case of within-group designs). The rationale for each of these three features is detailed elsewhere.[15]

When any of the three key features is missing, as often is the case when attempting to assess causal propositions in the real world rather than in the lab—for example, if we want to test the proposition that the assassination of a sitting United States president causes the stock market to decline, we cannot have some United States presidents randomly assassinated and others not—we have what is termed a "quasi-experiment." Although not as "clean" as experiments, quasi-experiments also enable one to draw causal inferences[16] and important treatises have been written on how this may be accomplished.[17] Elucidation of these designs (having names such as "untreated control group designs," "cohort designs," "regression continuity designs" and "interrupted time series designs") requires extensive discussion and is also beyond the present scope of this article. However, the subject of quasi-experimental designs (which include designs having no control group) warrant mention if only to sensitize both courts and counsel to their existence, applicability and value.

---

15. See Chapter 8 in Jacob Jacoby, The Role of the Judiciary in Creating and Fostering Junk Science, Forthcoming.

16. To place quasi-experimentation in perspective, recognize that *none* of the three distinguishing features of experiments may be present and scientists may still be able to draw valid causal inferences. Astronomy has made incredible progress in arriving at valid causal inferences without relying upon experimental designs.

17. See Donald T. Campbell and Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research (1966) (hereinafter Campbell & Stanley, Designs for Research); Thomas D. Cook and Donald T. Campbell, Quasi-Experimentation: Design and Analysis Issues for Field Settings (1979) (hereinafter Cook & Campbell, Issues for Field Settings); Thomas D. Cook, Donald T. Campbell and Laura Peracchio, Quasi-Experimentation, in The Handbook of Industrial and Organizational Psychology 491-576 (Marvin D. Dunnette and Leaetta M. Hough eds., second ed., 1990) (hereinafter Cook et al., Quasi-Experimentation).

IV. ARE CONTROLS ALWAYS REQUIRED FOR INTELLECTUAL PROPERTY SURVEYS?

Not all trademark, trade dress or advertising surveys demand the use of controls. At least three such situations can be identified.

A. When Assessing the Presence *of a State, Rather Than the* Cause *of the State*

In many instances, trademark surveys are directed to determining whether a particular mental state exists, not what caused its existence. As examples, we may want to know whether a particular term is or is not generic (or famous), not what caused it to be so. We may want to know whether or not a particular logo or slogan has secondary meaning, acquired distinctiveness or is famous, not what caused this to be so. These questions are analogous to a pollster seeking to determine what percent of the population intends to vote for a particular candidate (and being able to predict therefrom with a reasonable degree of certainty the outcome of the election), but not be able to provide any insight on what caused these intentions. Under such circumstances, depending upon whatever else must be considered or addressed, there may be no need for controls.[18]

B. When Assessing Whether Something Is Not *Causal*

If the issue to be determined is whether something alleged to be causal is in fact not causal, depending upon whatever other factors must be considered or addressed, there may be no need for controls. For example, plaintiffs in The Novus Group, Inc. v. Dean Witter, Discover & Co. Novus Credit Services, Inc. and Discover Card Services, Inc.[19] were in the business of selling tickets to the circus and the Ice Capades over the Internet. Defendant's Discover credit card and promotional materials (including decals on participating merchant doors, windows and cash registers) contain the phrase "Novus." As first to use the term "Novus" in commerce, plaintiffs alleged that when consumers familiar with defendant's card and materials used plaintiff's services, it would cause reverse confusion.

---

18. While not necessarily applicable for assessing causation, controls may still be useful for other purposes. For example, consider a multiple choice question such as: "Which of the following currently is a member of the U.S. Supreme Court: Richard Posner, Sylvester Stallone, Ruth Bader Ginsburg or Felix Frankfurter?" In the present instance, while not used in the service of assessing a causal question, the three incorrect names serve as controls used to assess guessing.

19. No. CIV.A. 93-1056-A (E.D. Vir. April 1, 1994).

The experimental stimulus I devised for that matter—a professionally developed eight-minute videotape of the dozen or so video screens a consumer using plaintiff's services would have to go through to make a purchase—was quite costly. When the findings revealed less than 2 percent confusion, it was decided to forego developing a corresponding control video to use with a control group. In pre-trial papers, plaintiff's expert claimed the study was flawed because, although it addressed a causal proposition, it contained no control group. At trial, I testified that the situation was analogous to testing a drug claimed to cure cancer. If tests of the drug revealed no one was cured, there would be no need for a placebo control group. Understanding this logic, the jury held for defendant.

It should be recognized that reliance on a single experimental group to determine non-causation entails certain risks. Had the test revealed not 2 percent, but a 20 percent level of confusion, defendant might have had a problem. Although it would have been possible to develop a control videotape and have it tested with a control group (and then subtract the level of confusion found with the Control group from the level found with the Experimental group to arrive at an estimate of "net" confusion), the fact that subjects would not have been randomly assigned to the Experimental and Control groups could have posed a problem, especially if some event (e.g., news coverage given to the dispute) intervened between testing of the Experimental group and testing of the Control group.

C. When Alternative Explanations Are Few and Can Be Ruled Out by Other Means

There are instances where, without the benefit of either a pre-test or an independent control group, the "one-group, post-test only" design is capable of determining causation. For example, entomologists often are able to identify which species of insect decimated trees by the tell-tale signature borings left by that species. Similar examples come from criminal law.[20] This approach relies on the presumption that certain causes leave unique signatures on the effect so that, given the effect, one can identify the cause. As discussed by seminal thinkers on experimental design, the "one-group, post-test only" design (see Design 1 described earlier) is capable of leading to reasonable inferences regarding causation under conditions where "the effect has to stand out, the pattern of evidence surrounding it has to be clear, the potential causes all have to be known, and auxiliary

20. M. Scriven, Maximizing the Power of Causal Investigations: The Modus Operandi Method, in Evaluation Studies Review Annual (R.F. Conner, D.G. Altman & C. Jackson eds. 1984).

information has to be available for discriminating among alternatives when several are available."[21] An example of how this applies in the trademark arena surfaces in National Football League Properties and Green Bay Packers v. ProStyle,[22] a matter discussed in a subsequent section of this article.

V. EXAMPLES OF EFFECTIVE AND INEFFECTIVE USE OF EXPERIMENTAL DESIGNS

Before discussing a number of basic considerations involved in selecting and/or devising Controls, several examples of intellectual property surveys are provided to illustrate how the basic designs have been used and sometimes misused or misinterpreted in trademark and deceptive advertising matters. Throughout this discussion, bear the following in mind. The most important percentage is not the percent confusion found with the allegedly confusing test item, nor the percent confusion found with the control item. Rather, it is the difference between the two or, in the parlance of the field, the "net confusion" that remains after subtracting the percent of confusion obtained with the Control stimulus (which is assumed to represent an amalgam of different forms of "noise") from the percent of confusion found with the Experimental stimulus. As employed in trademark litigation surveys, the formula is basic: Net effect due to the putative cause = [level of effect found in the test group] minus [level of effect found with the control group].

A. National Football League Properties, Inc. v. Wichita Falls Sportswear, Inc.[23]

One of the first uses of experimental design in a trademark survey was plaintiff's study in NFL Properties v. Wichita Falls. In that matter, the court was considering whether to permit defendants to manufacture and sell unauthorized football replica jerseys bearing the distinctive colors and names of National Football League teams, players and cities—providing each garment bore a sewn-in label stating, "not authorized or sponsored by the NFL." The question was whether these labels would be sufficient to reduce confusion to minimal, non-actionable levels.

---

21. Cook et al., Quasi-Experimentation, supra n.17 at 517.

22. National Football League Properties, Inc. and Green Bay Packers, Inc. v. ProStyle, Inc. and Sheri Tanner, No. 96-C-1404 (E.D. Wis. 1997); National Football League Properties, Inc. and Green Bay Packers, Inc. v. ProStyle, Inc. and Sheri Tanner, 16 F. Supp. 2d 1012 (E.D. Wis. 1998); National Football League Properties, Inc. v. ProStyle, Inc., 57 F. Supp. 2d 665 (E.D. Wis. 1999). As noted infra, the author was the expert who prepared the surveys discussed in these opinions.

23. 532 F. Supp. 651, 215 U.S.P.Q. 175 (W.D. Wash. 1982).

Although involving considerably greater detail (e.g., it was a national probability survey involving more than 3,700 respondents), at core, the survey I designed for plaintiff was a "between-group" design involving three Experimental and two Control groups. Subjects in the Experimental (E) groups were shown shirts bearing the proposed disclaimer label, while those in the Control (C) groups were shown the corresponding shirts without these labels. Using X to represent exposure to the disclaimer, the five groups were as follows:

Group		Assignment		
E1	City names (e.g., Pittsburgh)	R	X	O
E2	Player names (e.g., Terry Bradshaw)	R	X	O
E3	Team nicknames (e.g., Steelers)	R	X	O
C1	City names (e.g., Pittsburgh)	R		O
C2	Player names (e.g., Terry Bradshaw)	R		O

Time:--------1-------2----- →

"The results revealed . . . that 58% of the respondents shown a jersey without a disclaimer thought authorization was necessary whereas 59.1% of those shown the same jersey with a disclaimer believed authorization was necessary. Without qualification then, the disclaimer exerted . . . no corrective impact on [reducing] confusion [to a non-actionable level]."[24] Indicating it was placing great reliance upon these findings, the court held for plaintiff.

B. Schering Corporation v. Schering Aktiengesellschaft and Berlex Laboratories[25]

In a number of instances, the researcher is confronted with the need for more than one Control group. Consider the case of Schering, the United States pharmaceutical company, against Schering AG, the giant German pharmaceutical firm and its United States subsidiary Berlex Laboratories, a New Jersey firm that markets prescription pharmaceuticals known as "X-ray contrast media." Manufactured by Schering AG and sold through Berlex, these drugs are used in hospitals and diagnostic centers throughout the United States. Though Schering AG is a world leader in the manufacture and sale of numerous pharmaceutical products, at the time the suit was filed, Berlex did not market any other products in the United States manufactured by Schering AG.

---

24. Jacob Jacoby and Robert L. Raskopf, Disclaimers in Trademark Infringement Litigation: More Trouble Than They Are Worth? 76 TMR 35, 53 (1986). Greater detail on this study is provided in that article.

25. 667 F. Supp. 175 (D.N.J. 1987).

However, defendants contended that if they chose to do so, they should be permitted to promote said products in conjunction with the phrase "Schering AG, West Germany."

At the time the suit was filed, it appeared that Berlex would be expanding its product line under the Schering AG name with products that would be promoted to physicians and dispensed by pharmacists. To prevent these physicians and pharmacists who, in turn, exert influence over lay consumers, from being misled, Berlex intended to insert a disclaimer in its advertising, on its packages and in its detail materials much like the one it had earlier used in other promotional materials distributed in Europe. This disclaimer would read: "Schering AG, West Germany, is not connected with Schering-Plough Corporation or Schering Corporation, Kenilworth, New Jersey."

In designing the research for this matter, I used a modified version of Design 5. A total of 300 physicians and 300 pharmacists were randomly assigned to one of three groups, either one of two Experimental groups or to a Control group, resulting in 100 physicians and 100 pharmacists per group. To simplify description, attention here is confined to the study involving the 300 physicians.

The purpose of the first Experimental group was to determine whether exposure to the disclaimer (designated X_D below) would produce the intended effect, namely, dispel confusion. Since there was the possibility that some physicians might already know or guess of the independence of the two Scherings even without exposure to the disclaimer, gauging the impact of the disclaimer required the use of a "no-exposure" (no disclaimer) Control group—the basic "two-group" situation depicted in Design 5 outlined earlier.

However, there was reason to believe that exposure to a disclaimer might actually increase rather than decrease the chances of extracting an erroneous message. Given considerable evidence showing that consumers often do not read all the verbiage in an ad or on a package (especially the fine print), this could occur if the respondent scanned or retained only that part of the disclaimer mentioning the key name(s), thereby either neglecting or forgetting the part that spoke of there being "no" connection between the two. This is considered especially likely to occur when the negator is a single, small, easily overlooked word such as "no." If such was the case, it would mean that consumers would misperceive the disclaimer, interpreting it to be a "claimer" of affiliation and/or authorization. Thus, if Berlex were to use its proposed disclaimer, there was the chance it might increase rather than decrease confusion.

To test this proposition, respondents in a second Experimental group were exposed to a "modified disclaimer" (designated below

as X_{MD}). This was created by removing the word "not" from the original disclaimer, thereby effectively creating a "claimer," essentially a statement of affiliation rather than one of non-affiliation. That statement was: "Schering AG, West Germany is connected with Schering-Plough Corporation or Schering Corporation, Kenilworth, New Jersey." The design can thus be depicted as below.

Group		Assignment		
E1	Disclaimer	R	X_D	O
E2	Modified Disclaimer (the "claimer")	R	X_{MD}	O
C	No disclaimer/claimer	R		O

Time:----------1-------2----- →

As it turned out, the findings showed that both the disclaimer and modified disclaimer were comparably ineffective in reducing the likelihood of confusion. All three conditions resulted in approximately the same levels of residual confusion. The clear implication of the research was that consumers were unlikely to attend to any form of disclaimer or claimer, even when each is repeated more than a dozen times, as was the case in the brochures used as the test stimuli. As the court commented, the survey provided "very persuasive evidence of the tendency to abbreviate the Schering names and persuasive evidence of confusion."[26]

C. The Gillette Co. v. Wilkinson Sword, Inc. and Friedman Benjamin, Inc.[27]

Both of the experiments described above were "between-group" designs, that is, they relied on separate Experimental and Control groups. The controls were thus "external" to the Experimental group. In some circumstances, it is possible to design a "within-group" study, one where each subject is able to serve as his or her own control. Examples of experiments using "internal" controls are the three surveys I conducted for plaintiff in Gillette v. Wilkinson Sword.

Plaintiff in that matter alleged there were deceptive superiority claims in defendant's advertisements and packaging for its razor blades. Three studies were conducted to determine what implied message the defendant's commercials conveyed to consumers. Each subject was shown either one of defendant's two

26. Id. at 189. For greater detail on the study, see Jacob Jacoby and George J. Szybillo, Why Disclaimers Fail, 84 TMR 224 (1994).

27. 1991 U.S. Dist. LEXIS 21006 (S.D.N.Y. Jan. 9, 1991). Docket #89CV3586 (KMW). Findings of Fact and Conclusions of Law, Hon. Kimba M. Wood.

TV commercials or its packaging, then asked a series of open-ended questions followed by several closed-ended questions. The court's description of this research explains how internal control questions are developed and used.

> Both open-ended and closed-ended questions that are properly constructed represent reliable and valid methods by which to test consumer comprehension of advertising. Because answers to closed-ended questions can reflect some level of "noise" (i.e., misunderstanding that is attributable not to the advertisement but to the communication process in general), a "control" question was used in each of the three Jacoby communication studies to try to determine the level of "noise" present. In the control questions, respondents were asked whether they received one or more additional meanings from the communication. As in the closed-ended "test" questions, the response options included a meaning that the communication expressly contained (a correct answer), as well as one or more meanings that the communication did not contain (an obviously incorrect answer). The percentage of incorrect answers to a properly constructed control question can be used as a surrogate for misunderstanding inherent in the communication process. The level of misunderstanding not necessarily attributable to the advertiser can be eliminated from the results by subtracting the percentage of incorrect answers to the control question from the level of miscommunication found in the closed-ended test question. The court finds this is an appropriate procedure to assist in making the determination whether a not insubstantial number of consumers are being confused by something attributable to the advertiser.[28]

Such a within-group design employing internal controls may be depicted as follows:

Group		
One group	X	O
Time: ----------1----------2--------- →		

When developing internal control questions for within-group designs, care must be taken to insure that, regardless of whether the Experimental or Control question is presented first, we can be reasonably certain that there is little to no chance it will influence how the subject interprets and reacts to the question asked second. When this requirement does not constrain which questions, test or

28. Id. at *21-*22.

control, should be asked first, it seems preferable to randomize (or at least systematically rotate) their order across respondents.

D. Hershey Foods Corp. and Homestead, Inc. v. Mars, Inc.[29]

Questions are not the only elements of a study that can serve a within-group internal control function. The stimuli (products, packages, ads, etc.) to which respondents are exposed can serve this function as well. As an example, consider Hershey v. Mars. Concerned that defendant's use of the same shades of orange, yellow and brown colors on packages for M&M's Chocolate Covered Peanut Butter Candies that plaintiff used on REESE'S Peanut Butter Cups would cause a blurring of the distinctiveness of what plaintiff believed were its famous shades and combinations of these colors, I was asked by plaintiff to design and conduct two surveys. One sought to determine the fame or renown not of the entire trade dress of REESE'S Peanut Butter Cups (which, among other elements, contains a saw-toothed representation of a peanut butter cup), but simply the combination, proportions and juxtaposition of its three colors. The second survey sought to determine if the packaging for M&M's Chocolate Covered Peanut Butter Candies would be associated with REESE'S Peanut Butter Cups, thereby blurring the distinctiveness of this combination of colors for candies.

Precise color-faithful representations were made of the two packages at issue, as well as of the packages for six other authentic brands—SNICKERS, BUTTERFINGER, NESTLÉ CRUNCH, SKOR, GOLDENBERG'S PEANUT CHEWS and MILKA, the latter being a European brand not sold in the United States. To avoid having the size or shape of the various packages telegraph which brand was being depicted, each representation was provided as a 5.5 inch x 2.5 inch card. "The stimuli depicted their respective colors, as well as the proportion, arrangement and juxtaposition of those colors. . . . As a control, the two groups [described below] were also shown a representation of an 'inverted Snickers' package, a package with all the design elements of a Snickers but with different colors than the actual Snickers package."[30] To avoid having distinctive typescripts or logos telegraph which brand was being depicted: "In all of the representations, including the two test representations and all seven comparison representations, the designation 'Brand X' (in upper case slightly slanted block letters) was used in place of the

29. 998 F. Supp. 500 (M.D. Pa. 1998).

30. Id. at 510.

real word marks or brand names in the colors of the real words. . . ."[31]

Respondents were randomly assigned to one of two test groups, either the group used to test fame and renown, or the group used to test blurring. These may be depicted as follows:

Group	Assignment		
Fame/renown group	R	X	O
Blurring group	R	X	O
	Time----------1------2--------→		

The respondents in each group were asked questions regarding eight different representations. In addition to being asked about the REESE'S representation, respondents assigned to the "Fame/renown group" were asked questions regarding the representations for the six authentic brands as well as for the "inverted Snickers" control. In addition to being asked about the M&M's representation, respondents assigned to the "Fame/renown group" were asked questions regarding the representations for the six authentic brands as well as for the "inverted Snickers" control. Before being shown any cards, respondents were told: "[E]ach of these cards shows the packaging for a different brand of candy. Although we've disguised the brand and manufacturer names. . . . As I show you each card, I'd like you to tell me if you think you know which brand it is or is supposed to be." As the eight representations for each group essentially formed a deck of cards, before being shown any of these, the interviewer was required to shuffle the cards. This insured a random presentation order. Then, for each card, the respondent was asked "If you think you know, what brand of candy comes in this package?" If they answered with a name, they were then asked "What, in particular, makes you think it is ___?"

For the first group, 94 percent of the respondents correctly identified the REESE'S representation and almost all of these (88 percent of the entire group) referred to the colors as the basis for their identification. For the second group, while only 7 percent correctly identified the M&M's representation as M&M's, 51 percent misidentified the M&M's representation as being REESE'S

---

31. Id. In this regard, note that some descriptions of the study (Steven B. Pokotilow and Stephen A. Fefferman, FTDA Survey Evidence: Does Existing Case Law Provide Any Guidance for Constructing a Survey? 91 TMR 1150 (2001)) are in error when they assert that "the packages were marked Brand X in a script that resembled the Reese's [distinctive script] typestyle" (91 TMR at 1162). Defendant had argued that the REESE'S stimulus improperly depicted the designation Brand X in the same typestyle as on an authentic REESE'S package, but the court found that "the slant is simply not important" and the depiction did not use the same type style. 998 F. Supp. at 515. In point of fact, Brand X appeared in slanted, separated block letters, not script.

and then referred to the colors as the basis for this identification. Regarding the first group, the court stated "We accept for purposes of this motion that Dr. Jacoby's survey has established the renown of plaintiff's mark, defined by the plaintiff as its trade dress in the colors of the Reese's package in a certain juxtaposition. . . ."[32] However, the court rejected the findings from the second group as having probative value in regard to determining blurring or dilution, agreeing with defendant's argument that "the absences of visual clues on the M&M's stimulus like the cascading M&M's, with the presence of analogous clues on some of the other brand stimuli, could have misled respondents into concluding that the M&M's stimulus was a representation of the Reese's trade dress."[33]

Without providing further illustrations, suffice it to say that within-group designs relying upon internal controls are often employed in trademark surveys. To recognize that this is so, all the reader need do is think of surveys that involve product (or ad) arrays.

E. Federal Trade Commission v. Kraft, Inc.[34]

Simply incorporating one or more Control groups provides no guarantee that the findings will be meaningful, useful or valid. Consider the study commissioned by the FTC in its case against Kraft.

The Kraft campaign at issue consisted of several allegedly deceptive TV and print ads. Each slice of KRAFT SINGLES processed American cheese is made from 5 ounces of whole milk, which yields the equivalent of 21 percent of the Recommended Daily Allowance (RDA) of calcium. However, approximately 1.5 ounces of milk is lost during processing, so that each slice contains the equivalent of approximately 16 percent to 17 percent of the RDA. By having the factually accurate claim that each slice is made from 5 ounces of whole milk appear in ads that also mention calcium, the FTC argued consumers would be deceived into believing each processed slice contained an amount of calcium equivalent to the amount found in five ounces of whole milk.

The experimental design employed in the FTC's research was similar for both the broadcast and print communications. For the broadcast communications, separate groups of Experimental subjects were exposed to one of two allegedly deceptive TV commercials, and the results from these groups were compared to those obtained from a Control group exposed to another Kraft commercial that said nothing about milk or calcium and therefore

32. 998 F. Supp. at 521 (M.D. Pa. 1998).

33. Id. at 519.

34. 970 F.2d 311 (7th Cir. 1992). The author was a witness for defendant in this matter.

was not in dispute. Similarly, for the print ads, the results from a group of Experimental subjects exposed to an allegedly deceptive print ad were compared with those obtained from a group of Control subjects exposed to another Kraft print ad that said nothing about milk or calcium and therefore was not in dispute. There were 100 respondents randomly assigned to each of the three Experimental and two Control groups. Though exposed to different communications, all 500 subjects underwent the same protocol and answered the same questions. The sequence of activities, which is unlike any design described thus far, was as follows.

Group	Assignment				
E1 TV commercial #1	R	X_{TV1}	O_1	X_{TV1}	O_2
E2 TV commercial #2	R	X_{TV2}	O_1	X_{TV2}	O_2
E3 Print ad	R	X_{P1}	O_1	X_{P1}	O_2
C1 TV commercial	R	X_{TV3}	O_1	X_{TV3}	O_2
C2 Print ad	R	X_{P2}	O_1	X_{P2}	O_2
	Time:----------1--------2--------3-------4--- →				

At Time 1, the subjects in each group were shown their respective ads along with distracter ads for two other products. At Time 2, an assessment of communication impact was made by asking the subjects a series of 10 multi-part questions, including the following:

Q.3 Do you remember seeing an ad for KRAFT SINGLES?

Q.4 What point does the Kraft ad make about the product? (PROBE:) What else?

Q.4a Is there anything else about the Kraft ad that stands out in your mind? (PROBE:) Is there something else?

Q.5a Does the ad give you any reasons why you should buy KRAFT SINGLES?

Q.6 Does the ad say or suggest anything about the nutritional value of KRAFT SINGLES, or about how healthy or good they are for you?

Q.8 Does the ad say or suggest anything about the milk content of KRAFT SINGLES?

Q.9 (If yes to Q 8:) You said the ad mentioned the milk content of KRAFT SINGLES. What does the milk content of KRAFT SINGLES mean to you?

Q.10 Does the ad say or suggest anything about the calcium in KRAFT SINGLES?

Thus, either because they mentioned the word "calcium" themselves or because they were asked a series of increasingly

focused questions that culminated in a question about "calcium," the testing protocol guaranteed that, by the end of Time 2, all 500 respondents—including the 200 respondents in the two Control groups that had been exposed to communications saying nothing about either milk or calcium—had been "primed." That is, they now had the notion of calcium planted in their minds.

After the "priming" created by these questions, the third phase of the protocol (Time 3) involved returning to these subjects the Kraft communication they had reviewed at Time 1 and instructing them to review this communication "one more time." Immediately afterward, phase four (Time 4) involved asking these subjects the following questions:

> Q.11 Does this ad say or suggest anything about the amount of calcium in a slice of KRAFT SINGLES compared to the amount of calcium in five ounces of milk?
>
> Q.12 Does this ad compare KRAFT SINGLES to imitation cheese slices?
>
> Q.13 Does this ad make any direct comparisons between KRAFT SINGLES and other cheese slices?
>
> Q.14 Based on this ad, do you think KRAFT SINGLES have more calcium, the same amount of calcium, or less calcium than those cheese slices they are being compared to?

Based on their answers to these questions, particularly Question 14, the FTC's expert concluded his experiments demonstrated that the three contested ads had caused consumer deception. In contrast, examination of the data revealed it was the testing protocol, not the allegedly deceptive nature of the ads, that was the most likely cause for the spurious findings of likely deception. Specifically, as a result of the various skip patterns associated with the questions,[35] 91 of the 200 Control subjects and 246 of the 300 Experimental subjects ended up being asked Question 14. When you think about it, no Control subject should have reached that point, as the Control ads had been carefully selected to serve as controls for the basic reason that they conveyed no message about either milk or calcium. Thus, when asked Question 11, all the Control subjects should have replied, "The ad I saw doesn't say anything about calcium." Despite this, nearly half (91/200 = 45.5 percent) said that the Kraft ads did say something about calcium. When asked Question 14, nearly two-thirds (57/91 = 63 percent) of these Control group respondents replied, "The ad said that KRAFT SINGLES have more calcium." If one compares this 63 percent to the 74 percent (181 of the 246) of Experimental subjects who replied that "the ad said KRAFT

35. "Skip pattern" is a term of art and refers to instances where how a respondent answers one question determines which questions that respondent subsequently is (or is not) asked.

SINGLES have more calcium," then it becomes obvious—although not to those at the FTC—that it was something about the testing procedure and/or questions, rather than exposure to the allegedly deceptive ads, that caused the spurious finding of deception.[36] At the very least, this renders the findings uninterpretable and, therefore, unreliable.

In other words, simply including one or more Control groups does not insure meaningful findings. Indeed, their inclusion may be more decorative than anything else, serving no other function than that of pseudoscientific legerdemain—the stuff of smoke screens and mirrors rather than honest inquiry. Sometimes, courts are astute enough to recognize this, as described in the next case.

F. Cumberland Packing Corp. and Stadt Corp. v. Monsanto Company, The NutraSweet Company, et al.[37]

Spanning many pages,[38] the published opinion in this matter provides what is perhaps the most detailed discussion by a court as to why the alleged controls in plaintiff's three surveys (as well as many other features of these surveys) were inadequate, leading the court to conclude that all three "studies [were] flawed and the survey results untrustworthy and unreliable."[39]

Briefly, in a study plaintiff's researcher conducted for a preliminary injunction hearing, respondents were taken into one room, shown a box of NatraTaste sugar substitute, then taken to a second room where they were shown a box of NUTRASWEET along with the boxes of four other authentic brands, after which they were asked questions to assess confusion.[40] This within-group design employing internal controls may be depicted as follows:

Group			
One group		X	O
	Time:----------1--------2--------- →		

The court criticized use of these other brands as controls because none of the controls had the dominant color blue, where as the plaintiff's NatraTaste and defendant's NUTRASWEET products featured the color blue. An analysis of the data by the

36. For a detailed discussion of the many serious flaws inherent in the FTC study, see Jacob Jacoby and George J. Szybillo, Consumer Research in FTC Versus Kraft: A Case of Heads We Win, Tails You Lose? 14 Journal of Public Policy and Marketing 1-14 (1995).

37. 32 F. Supp. 2d 561 (E.D.N.Y. 1999). This writer provided an expert report to the court on behalf of defendant.

38. Id. at 570-79.

39. Id. at 571.

40. Id. at 574-75.

court "showed that the two most common reasons people gave for thinking NUTRASWEET was made by the same company that made NatraTaste was the overall blue coloring of the boxes and the similarities of the names."[41] Because the court had already ruled that the color blue served a functional purpose to identify aspartame sweeteners, confusion caused by this non-protectable aspect of the trade dress was not relevant.[42]

The researcher conducted another study employing four derived controls (i.e., controls developed by the experimenter for the purpose of the research, but which have no presence in the real-world of commerce). Among other elements, the package depicted a coffee-cup. Two of the derived controls left the coffee-cup on the package; two removed the coffee cup. Two of the controls replaced the multi-toned blue with a solid darker (albeit still quite similar) shade than the primary blue of the original box. Again, the court's observations are directly on point. In regard to the coffee cup and color blue, the court found these elements to be generic when viewed in isolation. Sweetener boxes commonly feature images of a coffee cup, glass of iced tea, or individually wrapped paper packets that do not help distinguish plaintiff's products from others. Furthermore, as consumers buying artificial sweeteners associate the color blue with aspartame sweeteners, the dominant coloring of these boxes serves a functional purpose and does not differentiate them from other products.

The court gave numerous reasons why the "within" controls in the first two studies, and the separate control study in its entirety, were inadequate. Among these, the most fundamental included: (1) In a trade dress case, the controls need to identify and parse out confusion due to non-trade dress factors. What the court at one point characterized as plaintiff's "so-called controls" failed to do so.[43] (2) It is inappropriate to accept blindly the percentages associated with a control without examining the reasons respondents give for their answers. In many instances, the answers to a follow-up "Why did you say that?" revealed that the confusion associated with the controls had nothing whatever to do with the trade dress issues before the court. (3) The separate control study had "no relevance to the . . . issue."[44]

Another study plaintiff conducted for this matter involved a between (i.e., external control) group design.

> Seventy-six people were shown an EQUAL packet with a picture of a coffee mug on it (Group I) and seventy-seven people were shown an EQUAL packet with a picture of a

41. Id. at 568.

42. Id. at 575.

43. Id. at 572.

44. Id. at 578.

> strawberry on it (Group II). Each person in Groups I and II were [then] asked, "'Which of these, if either, do you think contains more NutraSweet per packet?". . . Of the 76 people in Group I, 36 percent said they thought the NutraSweet packet contained more NutraSweet-brand aspartame, 32 percent said EQUAL, and 33 percent answered, "the same" or "don't know." In other words, each choice received about one-third of the total responses, the result one would expect if the responses had been made at random. . . .
>
> [For Group II, those shown a packet bearing a strawberry instead of a coffee mug, the corresponding results were] 55 percent said they thought the NutraSweet packet contained more NutraSweet, 31 percent said EQUAL, and 15 percent answered, "the same" or "don't know."
>
> * * *
>
> Holding everything else constant, including the name NutraSweet, when one group was shown a picture of a strawberry and the other group shown a picture of a coffee mug, the results changed dramatically. At the least, the results suggest that the name is not the only statistically significant variable influencing which product people think has more NutraSweet . . . [w]hat is clear is that without the picture of the strawberry, [plaintiff's expert] was unable to find any correlation between the name NutraSweet and people's perceptions of whether the NutraSweet or EQUAL brand contained more NutraSweet-brand aspartame per packet.[45]

In other words, regardless of whether one uses internal controls or external controls, they need to be meaningful and appropriate for testing the question(s) of fact at issue. As the court stated: "In a test of a causal proposition the *appropriate* use of controls is crucial."[46] Though not italicized in the original, the word *appropriate* is italicized here because of its overriding importance. Controls are not crucial, only appropriate controls are. There must be an appropriate rationale underlying the selection or creation of a control; otherwise, it simply may be a mindless inclusion having neither logical nor scientific legitimacy. Attention now turns to discussing some of the circumstances that make a control "appropriate" and not "mindless."

45. Id. at 583-84.

46. Id. at 574 (emphasis added).

VI. CONSIDERATIONS IN SELECTING AND/OR DEVISING "CONTROLS"

As courts are coming to understand, there is no universal control, no one-size-fits-all-situations. Controls must be appropriate to the situation. If not, they serve no meaningful function. This section is devoted to discussing a number of considerations that go into selecting or devising appropriate controls.

A. Assessing a Single Alleged Causal Element

If there is a typical likelihood of confusion matter, it probably is the situation where a single element of the allegedly infringing item (e.g., either the entire trademark or a component of the mark; the sound of the mark; etc.) is identified as the factor alleged likely to cause confusion. Under such circumstances, it tends to be easy, relatively speaking, to select or devise an appropriate control. If one exists in the real world, the researcher selects an item that is the same (or as close to being the same) as the allegedly confusing item in all respects, except for the fact that it does not contain the allegedly infringing element. When such a "natural control" is not available, although not always easily accomplished, one strives to generate a reasonable "derived control." Depending upon what is at issue, the control may seek to be semantically, acoustically or visually (e.g., graphically) comparable to the allegedly infringing element.

Consider the matter of the Indianapolis Colts and National Football League v. Speros, the Baltimore Colts and the Canadian Football League (CFL).[47] For thirty-one years, plaintiffs operated a professional football team in Baltimore, Maryland, under the name the "Baltimore Colts." The team used a horseshoe as its logo and its predominant uniform color was a shade of medium to light blue. In 1984, the team moved to Indianapolis, Indiana and began playing under the name "Indianapolis Colts," retaining both the horseshoe logo and blue color.

The Canadian Football League (CFL) is a completely independent entity comprised of Canadian football teams. In 1993, the CFL expanded into the United States with a team based in Sacramento, California. During 1994, the CFL expanded into Shreveport, Louisiana, Las Vegas, Nevada and Baltimore, Maryland, with the latter team adopting the names "Baltimore Colts" and "Baltimore CFL Colts." Prior to its first scheduled

47. Indianapolis Colts, National Football League Properties, Inc. and National Football League v. Metropolitan Baltimore Football Club, Limited Partnership, James L. Speros and Canadian Football League, 31 U.S.P.Q.2d 1801 (D.C. Ind., 1994); Indianapolis Colts, Inc. et al. v. Metropolitan Baltimore Football Club Limited Partnership, et al., 34 F.3d 410, 416 (7th Cir. 1994).

game, merchandise bearing these names was available for purchase at retail outlets in and around Baltimore. This merchandise included three different versions of a tee-shirt and one baseball-type cap. These items used a similar shade of blue to that of the NFL's "Colts," said either "Baltimore Colts" or "Baltimore CFL Colts" and, as a logo, contained a stylized horse's head.

Believing the names "Baltimore Colts" and "Baltimore CFL Colts" would cause confusion among the relevant public as to affiliation and authorization, plaintiffs counsel retained me to design and conduct a study to determine whether (and, if so, to what extent) there was likely to be such confusion. All three versions of defendants' tee-shirts were tested. Because plaintiffs could not preclude the CFL expansion team from using the name "Baltimore," the color blue, or a horse head logo, the only element at issue was defendant's use of the term "Colts" to describe a professional football team located in Baltimore, Maryland. Given these circumstances, what should serve as the control?

From a scientific perspective, the cleanest and most defensible control involved deleting the single allegedly infringing element (the term "Colts"), replacing it with an equivalent term compatible with the horse head logo,[48] and leaving everything else intact. This way, the level of confusion found, if any, with this derived control could be said to reflect "noise" (such as might be caused by the question wording, other aspects of the measuring instrument, the interviewer, the test protocol, the respondent, and other forms of "noise"). Subtracting the noise level obtained with the Control shirts from the corresponding level found with tee-shirts containing the allegedly confusing "Colts" term would yield a fair estimate of likely confusion arising from the use of "Colts" in this context.

---

48. As there was nothing about the type font or style, or the sound of the name that could be protected, this was not a matter involving either visual or acoustic equivalence. In this instance, the fact that the control had to be compatible with the horse head logo dictated that it be semantically equivalent. Note that it sometimes is possible to find a control term that possesses a reasonable degree of semantic, visual and acoustical similarity, as illustrated by a matter now in dispute, Giant Brands, Inc. and Giant of Maryland LLC v. Giant Eagle, Inc. and Phoenix Intangible Holding Co., U.S.D.C., D. MD. Civil Action No. AW 02 CV-320. Since at least 1936, plaintiffs have provided and advertised retail supermarket services under the trademarks "Giant Food" and "Giant." Operating in different states, defendants use the name Giant Eagle in connection with supermarket services. After defendants acquired a chain of supermarkets in plaintiffs' trading area and re-opened them as Giant Eagle supermarkets, plaintiffs sued. In a survey commissioned by plaintiffs, the control I developed involved substituting the word "Great" for the word "Giant" each place it appeared as part of the name Giant Eagle in a supermarket ad. Both Great and Giant begin with the letter G and end with the letter T. Both contain five letters, three of which—g, a and t—not only are held in common, but also appear in the same sequence. In addition to having some degree of phonetic overlap, the two terms also possess some semantic overlap, as both "great" and "giant" imply something large.

This approach is consistent with the Federal Judicial Center's Reference Manual on Scientific Evidence. "In designing a control group study, the expert should select a stimulus for the control group that shares as many characteristics with the control group as possible, with the key exception of the characteristic whose influence is being assessed."[49] Since the Baltimore CFL Colts' horse head logo never was at issue, there was no justification for modifying or removing it from the control garments. Indeed, had the horse head logo been modified or deleted, one could expect defendants' counsel to have argued that, since the NFL's Indianapolis (née Baltimore) Colts used a horseshoe logo, in the context of its dramatically different horse head logo, the name "Baltimore CFL Colts" would not be confusing. Hence, whatever confusion surfaced in plaintiff's survey was simply a function of plaintiff having separated the "Baltimore CFL Colts" name from its real-world horse head logo context. At that point, instead of having a survey that assisted it in its deliberations, the trier of fact would be left with having to resolve a fundamental methodological and interpretive ambiguity (which, in scientific parlance, is termed a "confound"[50]). Thus, the horse head logo (measuring 14" x 8") emblazoned on the Experimental group garments had to be retained on the corresponding Control garments as well.

More importantly, since there now would have been two features that differed between Test and Control garments (Colts v. another nickname and horse head logo v. another logo), replacing the horse head logo would have left no way to trace unambiguously the impact attributable solely to the word "Colts," the element at issue. In other words, to determine whether the name "Colts" (and nothing else) was responsible for causing confusion in this context, tight scientific assessment required that it be the only element replaced, i.e., that the horse head logo continue to appear on the control garments.[51] This scientifically-dictated requirement prevented using any non-horse related name—something later suggested by the appellate court.[52] To have done so only would

---

49. Diamond, Reference Guide on Survey Research, supra n.8 at 258.

50. See Kaye & Freedman, Reference Guide on Statistics, supra n.8 at Section II.C.2.

51. As Judge Richard Posner of the 7th Circuit writes:

> "Controlled" experiments suppress features of the environment that are deemed irrelevant, in order to isolate the effect of the variable under investigation. But the experimenter may err in the design of the experiment. One of the excluded features may be the real cause of the phenomenon being observed, and the independent variable that the experimenter wanted to test and that he found to have causal significance may just be a correlate of the omitted variable.

The Problems of Jurisprudence 65 (1990).

52. "We don't like the name 'Baltimore Horses,' as we have said, but we doubt whether a more attractive 'Baltimore' name, the 'Baltimore Leopards' for example, would have generated the same level of confusion that the 'Baltimore CFL Colts' did." Indianapolis Colts, Inc. v. Metro Baltimore Football Club Ltd. Partnership, 34 F. 3d 410, 416 (7th Cir.

have *reduced* the level of confusion obtained with the Control garments, thereby *increasing* the difference between it and the level of confusion obtained with defendant's garments. The net result would have been a higher estimate of likely confusion—which would have favored plaintiffs, not defendants. At that point, defendants would have been justified in claiming that the use of another name (e.g., Leopards) biased the results in the plaintiff's favor.

If Horses is, to use some of Judge Posner's words, "unappealing" or not "attractive," consider the other semantically equivalent equine terms that could have been employed in conjunction with a horse head logo, namely, Broncos, Chargers, Fillies, Mares, Ponies, Stallions, Steeds.[53] As they were already in use as the names of other National Football League teams, the terms Broncos and Chargers could not be used. Most would agree that fillies, mares and ponies were too effete to serve as names for a professional football team. By virtue of it having been the name of another now defunct professional football team ("Birmingham [Alabama] Stallions"), the term Stallions had been "contaminated." This left "horses" and "steeds." It might have come down to a coin flip were it not for the fact that, while both had nearly the same number of letters as "colts," like "colts," "horses" had an "o" as its second letter and ended with an "s," thereby being visually more similar to Colts. All things considered, given that the horse head logo had to be retained, the powerful term "Horses" seemed the best choice.

The bottom line: Using Control shirts and caps that substituted the term Horses where Colts appeared, the study found high levels of likely confusion and was relied upon by the district court in holding for plaintiffs.[54] Notwithstanding its misgivings regarding the term Horses, the Seventh Circuit upheld the district court "crediting the major findings of the Jacoby study and inferring from it and the other evidence in the record that defendants' use of the name 'Baltimore CFL Colts' whether for the

---

1994). Had the name Colts been replaced with Leopards, with the latter now appearing above the horse head logo, not only would this look bizarre, but respondents would have readily understood that the garment was not natural and unlikely to be found offered for sale to the public.

53. Mustangs might have been another possibility not thought of at that time.

54. The district court made the following statements about the survey: "This Court has accepted the plaintiffs' survey as being a helpful indicator of the degree of similarity of the parties' marks." 31 U.S.P.Q.2d at 1809. "This Court finds this survey helpful in assessing whether the marks are so similar in appearance and suggestion as to cause confusion even though defendants have opined differently. . . . The survey shows that there is confusion in any case." 31 U.S.P.Q.2d at 1808.

team or merchandise was likely to confuse a substantial number of consumers."[55]

B. Assessing a Combination of Alleged Causal Elements

1. Preamble

Having just discussed a ruling by Judge Posner and as foundation for the ensuing discussion, it is worth noting one of the many passages Judge Posner has written regarding the interplay between science and the judiciary:

> And even if all the judges up and down the line agree [with decisions emanating from higher courts], their decisions have much less intrinsic persuasiveness than unanimous scientific judgments have, because judges' methods of inquiry are so much feebler than scientists' methods. (Does anyone doubt, as Justice Robert Jackson once remarked, that if there were a court above the Supreme Court a large fraction of the Supreme Court's decisions would be reversed?) . . . A lawyer who loses a case in the Supreme Court, a judge who is reversed by the Court, a law professor commenting on the Court's latest (and let us say unanimous decision)—none of these is speaking nonsense, or even violating professional etiquette, if he says the decision is wrong. Our legal discourse is not so positivistic that one is forbidden to appeal to a "higher law" even after the oracles of the law have spoken. . . .[56]
>
> [A] carapace of falsity and pretense surrounds law and is obscuring the enterprise. It is time we got rid of it.[57]

Because it provides an excellent foundation for elucidating many important issues that need to be considered when developing controls for both trademark confusion and trademark dilution (qua

---

55. Indianapolis Colts, Inc. v. Metro Baltimore Football Club Ltd. Partnership, 34 F. 3d 410, 416 (7th Cir. 1994). The court also commented that "it is unfortunate and perhaps a bit tricky that the subsample of consumers likely to buy merchandise with a team name on it was not limited to consumers likely to buy merchandise with a *football* team's name on it. . . ." I agree; it was unfortunate. However, consumers likely to buy "merchandise with a football team's name on it" are not completely different from consumers likely to buy "merchandise with a team's name on it." Consumers "likely to buy merchandise with a team's name on it" evidence a predisposition toward buying "merchandise with a football team's name on it" under various circumstances, e.g., as when their favorite team appears in a play-off game or in the NFL's Superbowl. Regardless, this question was of minor import. Not only was it asked after all the substantive questions had been asked (and, hence, could not have influenced the findings obtained with the substantive questions in any way), the data obtained from this question were not relied on either in the report or during oral testimony.

56. Richard A. Posner, The Problems of Jurisprudence, 79-80 (1990).

57. Id. at 469.

blurring) matters, Parts VI.B.2. through VI.G. below focus on the opinions regarding scientific procedure espoused by one district court. I will be relying upon virtually "unanimous scientific judgments" in arguing that the district court was "wrong," not once, but many times. As observed by Judge Posner, rendering such criticisms does not violate "professional etiquette"; rather, it represents an effort to produce understanding in an arena where understanding sometimes is not easy to come by.

2. Discussion

While developing a control is often, relatively speaking, easy when there is only a single putative cause, how does one select or devise a control when the effect (say, confusion, or dilution via a blurring of distinctiveness) is alleged to be caused by a combination of several trademark and/or trade dress elements? This is precisely the situation that applied in National Football League Properties v. ProStyle.[58] Consider how the court itself framed the issue in its July 25, 1997, Decision and Order:

> Defendants, through ProStyle, have recently and without plaintiffs' consent commenced selling in interstate commerce merchandise, including shirts, sweatshirts, dresses, swimsuits, caps and jackets, bearing the designations "PACK," "GREEN BAY P," with a player's name and number, "GBP CENTRAL DIVISION CHAMPIONS" AND ""DIVISION CHAMPIONS GREEN BAY." Defendants' merchandise often display the Packers' team colors, which are dark green and yellow, or variations thereof. Certain articles of defendants' merchandise also bear football indicia, including football helmets, in the Packers team colors or variations thereof, which are displayed in conjunction with the aforementioned designations or with the names of various Packers' players and the respective numerals worn by those players. Defendants' products are often interspersed in the marketplace with products officially licensed by plaintiffs. Defendants have advertised their products in interstate commerce through a mail-order catalog. . . . In the catalog, pictures of defendants' products are interspersed throughout defendants' catalog with color photographs of team members of the Packers in team uniforms and team helmets.[59]

---

58. National Football League Properties, Inc. and Green Bay Packers, Inc v. ProStyle, Inc. and Sheri Tanner, No. 96-C-1404 (E.D. Wis. 1997). National Football League Properties, Inc. and Green Bay Packers, Inc. v. ProStyle, Inc. and Sheri Tanner, 16 F. Supp. 2d 1012 (E.D. Wis. 1998). National Football League Properties, Inc. v. ProStyle, Inc, 57 F. Supp. 2d 665 (E.D. Wis. 1999).

59. National Football League Properties, Inc. and Green Bay Packers, Inc. v. ProStyle, Inc. and Sheri Tanner, No. 96-C-1404 (E.D. Wis. 1997) at 6-7.

A likelihood of confusion survey was designed, conducted and proffered.[60] It involved showing eight comparable groups of respondents one of eight shirts.[61] Four shirts were defendants' "as sold" garments. The other four were "control" shirts essentially identical in all respects to the former, except for having an allegedly infringing element characteristic of defendants' shirts replaced with a non-infringing element on the control shirt. For example, whereas defendants' garments used the name "Green Bay," the corresponding control garments were the same in all respects, except for the fact that "Green Bay" was replaced with "Ellison Bay," the name of another bay in northern Wisconsin. After being shown one of the test or control shirts, the respondent was asked the following question:

1a. What, if anything, do you think of when you see this shirt?

The answer was recorded verbatim. Note that having been shown defendants' garments, the answers to this question can be used to directly assess the blurring of distinctiveness component of dilution (Lanham Act, Section 43(c)).[62] The answers may also be used to infer whether plaintiff's marks have acquired secondary meaning. Note that Question 1a is completely open-ended. Hence, respondents could have answered by saying anything. As any respondent who answered "Green Bay" might have meant the City of Green Bay or the Green Bay Packers, respondents giving such an answer were asked a follow-up question:

1b. Do you mean anything in particular by "Green Bay"? (Probe once with: *Anything else?*)

Question 1b also was completely open-ended and made no mention of the municipality, the Packers or the NFL.

To assess confusion as to sponsorship (Lanham Act, Section 43(a)), the interviewer continued with:

2a. Do you think that in order to put out this shirt, the company that put it out. . .

did need to get permission,

did not *need to get permission,*

60. Jacob Jacoby, The Extent to Which Green and Yellow, When Seen in the Context of Other Pertinent "Cues," Have Acquired Secondary Meaning and are Likely to Cause Consumer Confusion. April 1997.

61. Actually, there were ten groups in all. As defendants' "player" shirt and its corresponding control addressed a peripheral issue, our attention is concentrated on defendants' four Green Bay Packers shirts and their corresponding control shirts.

62. Section 43(c) of the Lanham Act says nothing about demonstrating damages to anything beyond consumers' mental associations. However, since Ringling Bros.-Barnum & Bailey, Combined Shows, Inc. v. Utah Div. Of Travel Devel, 170 F.3d 449, 50 U.S.P.Q.2d 1065 (4th Cir. 1999), disagreement has been manifested among the federal appellate courts as to whether other actual economic harm must be shown. Discussion of this issue is beyond our present scope.

or you have no thoughts about this?

Respondents who answered "did need to get permission" were then asked Question 2b (*"From whom did they need to get permission?"*) and Question 3 (*"What makes you say that the people who put out this shirt needed to get permission from ____? Anything else?"*).

Defendants subsequently filed a motion in limine to exclude both the survey and this author's expert opinions regarding consumer reaction to the defendants' merchandise. Five arguments were presented in support of this motion. After devoting several pages to considering each of defendants' arguments, the court rejected four, but accepted the argument that "the survey's confusion question improperly asked for a legal conclusion."[63] In reaching this conclusion, the ProStyle court relied on a district court ruling that reached the same conclusion.[64] In doing so, both the ProStyle and earlier decision ignored numerous cases where the "need to get" formulation was accepted and relied upon by district[65] and appellate courts, including those in their own

63. 16 F. Supp. 2d at 1016.

64. Novo-Nordisk of North America v. Eli Lilly and Company 1996 WL 497018 7 n.24 (S.D.N.Y.) ("[R]espondents were asked whether the maker of each of the pens named on the package 'had to give its permission or approval to the maker of Humulin for the use of the Humulin cartridge in' the pen. This question mistakenly asks respondents what they believe is the legal requirement (because of the use of the phrase 'had to'), rather than asking them merely whether they believed that the maker of the Humulin did receive authorization to use the names of the pens.").

65. Because it also involved a Southern District of New York matter heard two years earlier, and because it identifies and defers to Second Circuit case law regarding the "had to get" versus "did receive" formulation, consider the following from Schiefflin & Co. v. The Jack Company of Boca, Inc. et al., 850 F. Supp. 232, 247, 31 U.S.P.Q.2d 1865 (S.D.N.Y. 1994):

> Defendants' objection to the question "Do you think the company that makes or distributes the product I showed you had to get authorization—that is, permission—from anyone else to market the product?" as a "legal" question is ineffective. The question . . . [is] certainly a relevant question under this Circuit's caselaw.

In accepting and according considerable weight to plaintiff's survey, the court in NFLP, Inc. v. Wichita Falls Sportswear, Inc., 532 F. Supp. 651, 659 (W.D. Wash, 1982) wrote: "interviewees who saw the team name on the shirt . . . believed that the manufacturer was required to obtain authorization from the NFL or one of the member clubs in order to manufacture the jerseys." Thus, the court indicates its understanding that both the question, and the findings obtained using the question, focused on whether consumers thought permission was required (i.e., had to be obtained), not whether permission had been obtained. Other examples include: National Football League Properties, Inc. v. New Jersey Giants, Inc., 637 F. Supp. 507 (D.N.J. 1986); Schering Corporation v. Schering Aktiengesellschaft, 667 F. Supp. 175 (D.N.J. 1987); Ferrari S.p.A. Esercizio v. McBurnie, 11 U.S.P.Q.2d 1843 (S.D. Cal. 1989); Ferrari S.p.A. Esercizio v. Roberts, 944 F.2d 1235 (6th Cir. 1991); Ferrari Esercizio S.p.A. v. Roberts, 739 F. Supp. 1138 (E.D. Tenn. 1990); Smartfoods, Inc. v. Hunt-Wesson, Inc., No. 3:92-CV-2061-D (N.D. Tex. Dec. 30, 1992); P.T.C. Brands, Inc. v. Conwood Company L.P., 28 U.S.P.Q.2d 1895, (W.D. Ky. 1993); Indianapolis Colts et al v. Metropolitan Baltimore Football Club et al., 31 U.S.P.Q.2d 1801, aff'd, 34 F.3d 410 (7th Cir. 1994). The surveys just cited were all conducted by the present author. While surveys conducted by others using the "had to get" permission meaning have been criticized on other grounds, these courts have not been critical of the "had to get" formulation. Examples include: The Sports Authority, Inc. and Intelligent Sports, Inc. v. Abercrombie &

circuits.[66] Even more important than case law precedent is the fact that, except in those rare instances where the respondent was privy to contractual discussions between the parties, asking respondents whether permission or authorization had been sought or received amounts to asking for a guess. Since guesses are generally held to be non-probative and accepting same would amount to junk science, why would a researcher or court want to rely on questions that encouraged respondents to guess? Commenting on the argument that the "need to get" formulation is incorrect because it "allows for the consumer's misunderstanding of the law," Professor McCarthy points out "it is consumer perception that creates 'the law' of whether permission is needed."[67]

Although the survey had been excluded because of its conclusions regarding confusion, the court had not been asked, nor had it considered, how the data obtained via Questions 1a-1b could be used to draw conclusions regarding dilution. As Questions 1a-1b were asked not about plaintiffs' garments, but about defendants' garments, it was thought the court would understand that the answers were directly relevant to assessing the "blurring of distinctiveness" component of dilution. Further, as it was the first question asked after a respondent was exposed to defendants' garments, there was no possibility that respondents' answers to

---

Fitch, Inc., et al., 965 F. Supp. 925, 939, 42 U.S.P.Q.2d 1662 (E.D. Mich. 1997) ("Do you believe Abercrombie & Fitch needed permission from The Sports Authority to use 'original outdoor authority'?"); The Rock and Roll Hall of Fame Museum, Inc. et al. v. Gentile Productions, et al., 71 F. Supp 2d 755, 762 (N.D. Ohio 1999) ("From whom do you believe that they would need permission or authorization?"); Lord Simon Cairns et al. v. Franklin Mint Co., et al., 107 F. Supp. 2d 1212, 1219, 55 U.S.P.Q.2d 1711 (C.D. Cal. 2000) ("The survey consisted of several questions. The first asked: 'does the company or organization that is selling this Diana, Princess of Wales product need to get permission or approval of any other company or organization before it could offer it for sale, or not?'").

66. Insofar as the Second Circuit is concerned, see Home Box Office v. Showtime/The Movie Channel, Inc., 832 F.2d 1311, 1315-16, 4 U.S.P.Q.2d 1789 (2d Cir. 1987); Charles of the Ritz Group v. Quality King Distrib., Inc. and Deborah International, 832 F.2d 1317, 1324, 4 U.S.P.Q.2d 1778 (2d Cir. 1987). In both these decisions, the appellate courts cite and rely upon J. Jacoby & R.L. Raskopf, Disclaimers in Trademark Infringement Litigation: More Trouble Than They Are Worth? 76 TMR 35 (1986), as a basis for reversing the burden of proof in disclaimer cases. As these appellate courts must certainly have realized, the key question in the survey described in that article used the "need to get" formulation. With regard to the Seventh Circuit, consider Indianapolis Colts et al v. Metropolitan Baltimore Football Club et al., 31 U.S.P.Q.2d 1801, aff'd, 34 F.3d 410 (7th Cir. 1994). Having written "The consumers [in Jacoby's survey] were asked . . . [w]hether the team or league needed someone's permission to use this name, and if so whose" (34 F.3d at 415), Judge Posner was obviously aware he was accepting the "need to get" formulation when he concurred with the district judge "in crediting the major findings of the Jacoby study and inferring from it . . . that the defendants' use of the name 'Baltimore CFL Colts' was likely to confuse a substantial number of consumers" (34 F.3d at 416).

67. 5 J. Thomas McCarthy, McCarthy on Trademarks and Unfair Competition, § 32:175 at 32-264 (3d ed. 2000). See related discussion in 3 McCarthy, § 24:9. Personal correspondence between this author and Professor McCarthy reveals that he believes "need to get," not "did get," is the proper formulation.

Question 1a would have been contaminated by their having been asked any of the subsequent questions used to assess confusion. Accordingly, eliminating all data pertaining to secondary meaning and confusion, the initial report was revised so that it focussed on what the respondents' answers to Question 1a revealed regarding dilution.[68]

The court would have none of this. As stated in its 1999 Order:

> The court concludes that Jacoby's survey, even as edited to avoid the court's earlier criticisms of it, is seriously flawed. The main problem with the survey (as edited for the second report) is that it essentially asks only one question, "What, if anything, do you think of when you see this shirt?". . . without further probing, see 5 J. McCarthy, Trademarks and Unfair Competition, § 32: 176 (1999) ("Without further probing, such a question may well be meaningless and irrelevant."), and without showing any control shirt to any survey respondents or asking any control questions.[69]

Are these "problems," as the court seems to think, or are they simply a reflection of the court's inadequate understanding of scientific research methodology?

First, there is no requirement in science (and likely none in law) that, to yield reliable and valid findings, questionnaires (or witness examinations) need to consist of some minimum number of questions. It is not the quantity of questions asked that matters, but their quality and responsiveness to the issue at hand. It requires only a single question to learn someone's age, social security number, marital status, etc., and it may require only a single question of a witness to learn all that is needed. In answer to a prosecuting attorney's question "Did you, or did you not, strangle your wife because she was having an affair with Mr. Doe?" suppose the witness answers "Yes; and I have no regrets over doing so!" Do we need to ask any other questions to establish motive, guilt and lack of remorse? Although subsequent questions fleshed out the details, it required but a single question to learn that President Nixon had installed a system for surreptitious tape recording in the Oval Office. If the question is not leading or otherwise biased and yields answers that are responsive, there is no legitimacy to the criticism that research is somehow flawed if but a single question was used to elicit this information.

Second, when the answers given by the respondents are clear and responsive, no treatise on survey research procedure will be found suggesting a need for further probing. When data adduced

---

68. Jacob Jacoby, The Fame of the Green Bay Packers as Embodied in Its Registered Marks, and the Extent to Which ProStyle Merchandise Dilutes and Is Likely to Cause Confusion with These Marks. September 1998.

69. 57 F. Supp. 2d 665, 668 (E.D. Wis. 1999).

from a single unbiased question are completely responsive to the issue being investigated, there is no legitimacy to the criticism that use of a single question represents a flaw. Where the answers to a single question are complete and responsive, further probing becomes a form of badgering, adds to interview length and decreases respondents' motivation to answer subsequent questions. When he wrote "may well be meaningless"—as quoted by the ProStyle court—Professor McCarthy used language precisely and clearly. By no stretch of any known logic, scientific or otherwise, does "may well be" equate to "necessarily will be" meaningless.

Third and most important are the court's various comments bearing on the issue of scientific "controls." At core, these involved two issues. The first issue concerns whether controls used to assess confusion are appropriate for assessing dilution. Discussion of this issue is reserved for a later section entitled "Controls Useful for Assessing Confusion are Not Necessarily Suitable for Assessing Dilution." Second, those portions of the FJC's Reference Manual on Scientific Evidence that discuss assessing causal questions do not provide any guidance regarding what should be done when a combination of factors are alleged to cause either confusion or dilution via blurring. As noted, one of these guides states: "In designing a control group study, the expert should select a stimulus for the control group that shares as many characteristics with the experimental stimulus as possible, with the key exception of the characteristic whose influence is being assessed."[70] Recognize that the reference here is to a single "characteristic," not to a combination of "characteristics." That author goes on to state: "Nor should the control stimulus share with the experimental stimulus the feature whose impact is being assessed."[71] Again, recognize that the reference is to a single "feature," not a combination of "features." Most importantly, recognize that were one to follow these guidelines, *any* control selected for use in the present matter would have *none* of the features characteristic of defendant's garments.

Plaintiffs in NFL v. ProStyle alleged both confusion and/or dilution via blurring would be caused by some *combination* of trademark and trade dress elements when these appeared on a football replica jersey or sweatshirt. In the case of a tee-shirt made to look like a football replica jersey, all the elements at issue would be present when that garment also contained: (1) a portion of the Green Bay Packers team name and/or a variation thereof ("Green Bay," "Green Bay P," or "Pack"), (2) dark green and yellow colors comparable to or suggestive of those used by the NFL's Green Bay

70. Diamond, Reference Guide on Survey Research, supra n.8 at 258.

71. Id.

Packers, (3) indicia designed to evoke an association to football (e.g., a football logo, a football helmet, goal posts, a gridiron, large player numerals on the back), (4) the name of a presently active Green Bay Packers player, and (5) designations such as "GBP Central Division Champions" and "Division Champions Green Bay." Under such circumstances—where deleting one allegedly confusing and/or diluting element would leave many others intact—how does one design a "fair" and proper control?

This complex issue—how to design a "fair" and proper control when deleting one allegedly confusing (or diluting) element would leave many others intact—provides the basis for much of the remaining discussion. As the discussion sometimes refers to confusion, sometimes to dilution qua blurring, and sometimes to both, it is necessary to be attentive to which of these is being discussed.

In a concurring opinion to Kumho, Justices Scalia, O'Connor and Thomas emphasized that a trial judge's discretion to admit or reject expert testimony "is not discretion to perform the function inadequately."[72] Great authority carries with it great responsibility. To avoid accepting "junk science" or rejecting quality science, when serving as a gatekeeper, the judge's appraisal needs to be informed. Otherwise, it cannot honestly or adequately perform its assigned function. However, as acknowledged by Justice Breyer, the author of Kumho, this task can be difficult, given that "judges are not scientists and do not have the scientific training that can facilitate the making of such decisions."[73]

C. Strong Controls v. Weak Controls

Although this point has not yet been discussed, controls can be selected or devised that vary in stringency. Whether designing a study for plaintiff or defendant, it is preferable to use strong controls—ones that, if they do cut against one of the parties, do so against the party proffering the survey. Under these circumstances, if the findings end up favoring the proponent of the survey, one can be more confident in the validity of the results. To see that such was the case in ProStyle, understand that, as always, there exists a universe of potential controls, often in the order of hundreds or thousands when the many possibilities and variations for derived controls are included. Clearly, no litigant would have the time, money and other resources to conduct

72. Kumho Tire, 119 S. Ct. at 1179.

73. General Electric Co. v. Joiner, 118 S. Ct. 512, 520 (1997). "[M]ost judges lack the scientific training that might facilitate the evaluation of scientific claims or the evaluation of expert witnesses who make such claims." Justice Stephen Breyer, Introduction, in Reference Manual on Scientific Evidence 4 (Federal Judicial Center ed., 2d ed. 2000).

empirical research using all the controls within this universe of possibilities. The researcher thus is forced to be selective and use judgment in selecting or devising the proper control(s). Whether consciously or not, the researcher engages in a series of "thought experiments," namely, "arguments concerning particular events or states of affairs of a hypothetical ... nature which lead to conclusions about the nature of the world around us."[74] The following example describes this process and illustrates various considerations that go into designing and selecting controls.

Suppose that, from the universe of possible controls, nine shirts had been selected for further consideration. These are designated as numbers 1 through 9 below. For purposes of comparison, three additional shirts are also described. Shirts #10 and #11 represent two of defendants' garments while Shirt #12 represents one of plaintiffs' authorized garments.

Shirt	*Type*	*Color(s)*	*Indicia on front*	*Other indicia*
1	Dress	white	none	none
2	Tee	white	none	none
3	Tee	white	"Ponies"	"Mumford H.S." + numerals on back
4	Tee	green & white	"New York"	"Jets" + numerals on back
5	Tee	white & blue	football logo	"Ellison Bay" + numerals on back
6	Tee	brown & yellow	"Ellison Bay"	Football helmet logo on front + numerals on back
7	Tee	green & yellow	"Ellison Bay"	Football helmet logo on front + numerals on back
8	Tee	green & yellow	"Groin Bay"	Football helmet logo on front + numerals on back
9	Tee	green & yellow	"Groan Bay"	Football helmet logo on front + numerals on back
10	Tee	green & yellow	"Green Bay"	Football helmet

74. Andrew D. Irvine, Thought Experiments in Scientific Reasoning, in Thought Experiments in Science and Philosophy 149-65 (Tamara Horowitz and Gerald J. Massey eds., 1991). See also Ernst Mach, On Thought Experiments, in In Knowledge and Error 134-47 (Thomas J. McCormack and Paul Foulkes trans., 1976) (1897); Albert Einstein, Relativity 25-27 (Robert W. Lawson trans., 15th ed. 1979) (1917).

Shirt	*Type*	*Color(s)*	*Indicia on front*	*Other indicia*
				logo on front + numerals on back
11	Tee	green & yellow	large "P"	"Green Bay" + numerals on back
12	Tee	green & yellow	"Packers"	"Green Bay" + numerals on back

Because Shirts #1 through #9 all qualify as shirts—not breakfast cereals, beers, automobiles or anything else—some might hold that any could serve as an acceptable control. Before evaluating each possible control, consider the following issues.

D. The Problem of Controls That Are "Too Similar"

When it comes to allegedly infringing merchandise, it is generally understood that the closer a second comer's mark or combination of marks, dress and other indicia come to those of a first comer, the greater is the likelihood that the second comer's use of these marks would cause confusion and, hence, be actionable. In similar fashion for famous marks, the closer the allegedly diluting item's source-identifying marks or other indicia come to plaintiff's famous mark(s) and indicia, the greater the likelihood that such an item would cause a blurring of distinctiveness resulting in dilution.

The same logic applies to selecting and developing controls. One can devise controls so that they have none, few or many characteristics in common with plaintiff's product and/or defendant's product. As one moves along the continuum toward having the control take on greater resemblance to plaintiff's product, at a certain point, the control would begin to generate confusion, perhaps rising to a level where the control itself would be actionable. Although "the expert should select a stimulus for the control group that shares as many characteristics with the control group as possible,"[75] it is generally understood that it makes no sense to use as a control a third party's product if that product would itself be actionable. Hence, when devising strong controls—controls that, when subtracted from the corresponding values obtained with the allegedly infringing test item, can be expected to provide the smallest and, therefore, most conservative estimate of likely confusion—one challenge is to devise a control that does not cross the threshold separating what would be non-actionable from what would be actionable. In likelihood of confusion and deceptive advertising surveys, many courts have used 15 percent as an approximate threshold, with percentages of net confusion or

75. Diamond, Reference Guide on Survey Research, supra n.8 at 258.

deception falling below this level not considered actionable (unless they involve matters of health or safety) and those above generally deemed actionable. Inasmuch as the percentages empirically derived from sample surveys are estimates of the true universe value and have a plus-or-minus range around them, in the best of all possible worlds, it would not be desirable for a control to yield confusion estimates that exceeded 10 percent. If it did, the control itself would begin to reach an actionable level of confusion and its utility as a control thereby compromised.[76] With this in mind, the nine potential controls selected for further consideration can now be examined.

E. Evaluating "Controls" from a Universe of Possibilities: An Illustration

Shirt #1 is a plain white dress shirt having one left breast pocket, long sleeves with standard single button cuffs and a button down collar. Although it qualifies as a "shirt," it looks completely different from a tee-shirt and is generally worn for a different purpose. Upon being shown Shirt #1 and asked, "*What, if anything, do you think of when you see this shirt?*" it could be reasonably predicted that no consumer would be likely to associate this plain white shirt with the Green Bay Packers. After all, even the most die-hard fan does not walk around with the Green Bay Packers uppermost in his mind at all times so that, upon being shown any object (e.g., a dress shirt, a toothbrush, a coffee cup, a candy bar, etc.) and asked, "*What, if anything, do you think of when you see this* ____?" he is likely to answer, "the Green Bay Packers." Thus, when the "zero percent association" to Green Bay Packers evoked by Shirt #1 is subtracted from the "percent association" found with the allegedly infringing Test shirt, the "net" association would remain at a maximum. Although technically it satisfies the definition of a Control, it is unlikely that such a weak control would be able to tease out anything but the most trivial type of "noise" or other error.

Now consider Shirt #2, a plain white tee-shirt having no graphic design elements whatever, and Shirt #3, a white tee-shirt containing the nickname "Ponies" on the front and the name Mumford H.S. on the back, but also having no design features that would make it look like a football replica jersey. Upon being shown either Shirt #2 or #3 and asked, "*What, if anything, do you think of when you see this shirt?*" the likelihood that respondents would answer, "Green Bay Packers," is so small that, for all practical

76. Notwithstanding the question of desirability, there are times when a control will yield estimates of confusion (or deception) that exceed 10 percent. Thus, it is important not to lose sight of the fact that, the question of desirability aside, the more important consideration is whether, when the control estimate is subtracted from the test estimate, the "net" exceeds 15 percent (or whatever value the court believes is appropriate).

purposes, it can be assumed to be zero. In terms of the guidance provided in the FJC's Reference Guide on Survey Research ("Nor should the control stimulus share with the experimental stimulus the feature whose impact is being assessed"[77]), Shirts #2 and #3 would qualify as appropriate controls, because they share no contested elements in common with defendants' infringing shirts (#10 and #11). However, they would be weak controls likely to yield inflated estimates of "net" confusion.

Shirt #4 is an NFL-authorized football replica jersey representing the New York Jets. This shirt—which has the appeal and advantage of being a real-world "natural" control, not a "derived" control—also has various elements in common with defendants' shirts. It uses a similar green color, displays a geographical name that consists of two monosyllabic terms and contains large football numerals on the back.[78] Moreover, it is likely to trigger among many respondents thoughts of football generally, and the National Football League specifically which, if anything, should increase the tendency for respondents to think of the Green Bay Packers. Suppose respondents shown Shirt #4 were asked, "*What, if anything, do you think of when you see this shirt?*" If some were prompted to say that looking at Shirt #4 made them think of the Green Bay Packers, would this be taken as evidence of dilution either by a court or the Green Bay Packers? Given that Shirt #4 possesses so few of the elements considered to represent the Green Bay Packers, of course not. Suppose some respondents shown Shirt #4 and asked, "*Do you think that in order to put out this shirt, the company that put it out did need to get permission, did* not *need to get permission, or you have no thoughts about this?*" then answered, "*did need to get permission,*" causing them to be asked, *"From whom did they need to get permission?"* The likelihood of respondents answering, "Green Bay Packers," to this "from whom?" question remains negligible, probably close to zero. Although more reasonable as a control than Shirts #1, #2 and #3, despite having the appeal and advantage of real-world authenticity, Shirt #4 also would be a weak control.

Designed to look like a football replica jersey, Shirt #5 moves closer to possessing elements characteristic of defendants' shirts. Although it contains neither green nor yellow, it does contain a football logo on the front, large player numerals on the back, and bears the name "Ellison Bay"—a two-word geographical name that has the word "Bay" in common with the name "Green Bay" on

77. Diamond, Reference Guide on Survey Research, supra n.8 at 258.

78. Because it also contains the word "bay," the Tampa Bay Buccaneers might have been used for illustrative purposes as well. However, although it contains the word "bay," the name Tampa contains two syllables and, other than generic black or white and unlike the green of the New York Jets, the team does not use any colors in common with the Green Bay Packers.

defendants' shirts. (Like Green Bay, Ellison Bay is a body of water off northern Wisconsin.) Suppose respondents shown Shirt #5 were asked, "*What, if anything, do you think of when you see this shirt*?" Given the few similarities between Shirt #5 and authorized Green Bay Packers merchandise, some might be prompted to think of the Green Bay Packers. On the other hand, given the many dissimilarities—the fact that Shirt #5 was white and blue, not green and yellow, did not use the word "Green"—the number of such individuals likely would be small. Regardless, because it possessed several features in common with defendants' shirts (Shirts # 10 and #11), it could be argued that Shirt #5 would be a reasonably stringent control.

Consider now Shirt #6. Designed to look like a football replica jersey, Shirt #6 contains a football helmet logo, one of the Packers team colors (yellow) and a geographical name consisting of two words, "Ellison Bay," the second of which is the same word as appears in the name "Green Bay." Because it contains more elements in common with defendants' shirts than does Shirt #5, Shirt #6 represents a more stringent control than does Shirt #5. For this reason, many would hold that Shirt #6 was well-suited to serve as a control for assessing likely confusion of defendants' merchandise.

Consider now Shirt #7. Except for substituting the single word "Ellison" for "Green," Shirt #7 is identical in *all* other respects to one of defendants' garments (#10). Both use precisely the same graphics and designs; both use the same shades of green and yellow found on Green Bay Packers merchandise; both use the same yellow football helmet; both use a name consisting of two words, the second of which is "Bay"; both bear the same large player numerals on back; etc. The only difference is that one shirt uses the word "Green" in its name while the other uses the word "Ellison." These distinctions may be summarized via the following nine-point comparison:[79]

79. It should be recognized that the defendants' and corresponding control shirts also were identical in many other respects. For example, instead of having different neck treatments (circular, v-neck or oval), both had circular necks. Instead of being made of different types of fabric, both were 100% cotton. Instead of being fabricated so that one appeared solid and another had a perforated mesh "tear-away" appearance, both appeared solid. Instead of one containing a single numeral while the other consisted of two or three numerals, both shirts consisted of the same two-digit numeral. Instead of one shirt using one style or font for this two-digit numeral and the other using a different style or font, both shirts used the same style/font.

Defendant's garment #10	*Control garment #7*
designed to look like football jersey	uses identical football jersey design
uses green as primary color	uses identical green as primary color
uses yellow as secondary color	uses identical yellow as secondary color
uses no other trim colors	uses no other trim colors
uses yellow football helmet	uses identical yellow football helmet
uses large player numeral on back	uses identical large player numeral on back
uses a two word name	uses a two word name
uses "Green" as first part of name	uses "Ellison" as first part of name
uses "Bay" as second part of name	uses "Bay" as second part of name

As confusion was alleged to derive from a combination of several allegedly infringing elements, to use Shirt #7 as a control when it contains all but one of these common elements renders Shirt #7 an exceptionally strong control—one likely to produce a conservative estimate of "net confusion." Since it was possible to use as controls shirts having a fewer number of elements in common (e.g., Shirt #6), some would say that, to a certain extent, using Shirt #7 as a control amounts to stacking the deck against plaintiff. (At the very least, some might contend that this could not and should not be used as a control because it does not satisfy the guidelines specified in the FJC's Reference Manual on Scientific Evidence, namely, that a control should *not* "share with the experimental stimulus the feature whose impact is being assessed."[80]) Yet this is precisely what was done; Shirt #7 corresponds to one of the shirts derived to serve as a "strong control" for assessing confusion in this matter.

As another way to understand why Shirt #7 represents a strong control, consider the following. There is widespread agreement within the relevant scientific community that, when interpreting incoming information from the outside world (such as when the individual seeks to identify an object encountered in the environment), human beings rarely attend to every single feature of the outside item. Instead, in the process of identifying this outside "stimulus," each and every one of us relies on a process

80. Diamond, Reference Guide on Survey Research, supra n.8 at 258.

called *"pattern recognition."*[81] This involves using some of the salient features of the outside object as a basis for probing the information stored in our memory about previously experienced objects in an effort to find one having features that seem to match the pattern of features we have apprehended from the object that is now the focus of our attention. When a sufficient number of features in the incoming information match the pattern of features associated with knowledge about an object in our memory, we tend to fill in the details and interpret the perceived object as an exemplar of the knowledge we have in memory.[82] As noted by Higgins,[83] when apprehending the outside world:

> Two basic variables influence the likelihood that some stored knowledge will be activated—[one of which is] the fit between the stored knowledge and the presented stimulus. . . . The greater the overlap between the features of some stored knowledge and the attended features of a stimulus, . . . the greater is the likelihood that the knowledge will be activated in the presence of the stimulus. . . .

That we do not need to—and, in fact, generally do not—pay attention to every single aspect of an external object before using what we have stored in our memory to interpret and identify that object becomes especially problematic in regard to designing controls to assess likely confusion when multiple elements are alleged to be responsible for causing confusion. While relying on pattern recognition usually results in correct identification and interpretation of the outside world, attending to a few salient features of the product may lead us to misinterpret and misidentify what we see. This is precisely the kind of phenomenon noted by Judge McKinney in Indianapolis Colts v. Metropolitan Baltimore Football Club:[84]

---

81. See, e.g., R. Reed Hunt & Henry C. Ellis, Fundamentals of Cognitive Psychology 51 (6th ed. 1999); Peter H. Lindsay and Donald A. Norman, Human Information Processing 75-80, 257-85 (1977). For a more extensive discussion prepared specifically for trademark practitioners, see Jacob Jacoby, The Psychological Foundations of Trademark Law. 91 TMR 1013, 1034-38 (2001).

82. Pattern recognition by humans is comparable to employing an Optical Character Recognition (OCR) program to scan and recognize the individual letters of a document. Just as the OCR software examines the features of each letter and seeks to match it with a known pattern already stored in the computer's memory, the mind seeks to match the features of the information incoming from the outside world with information already stored in memory. And just as the OCR reader may misidentify a particular item (for example, misreading a capital "I" as a lower case "l"), the human information processor will make comparable errors.

83. E.T. Higgins, Knowledge Activation: Accessibility, Applicability and Salience in Social Psychology: Handbook of Basic Principles 133, 135 (E.T. Higgins & A.W. Kruglanski eds., 1996).

84. 31 U.S.P.Q.2d 1801 (S.D. Ind. 1994).

Most people can probably recall a time when a return trip to the store was necessary because a word on a label triggered a memory that filled in the rest of the label, and caused them to select the wrong product. One's intention may be clear and still the wrong product was purchased.

Given this understanding of how human beings process information incoming from the outside world, it can be better appreciated why Shirt #7 represents a very stringent control. By having so many features in common with both plaintiffs' and defendants' shirt, it increases the likelihood that it, too, is likely to be confusing. In this way, when the estimate of likely confusion obtained with the control is subtracted from the estimate of likely confusion obtained with defendants' shirt, "net confusion" can only be decreased, thereby favoring defendant.

F. Controls Useful for Assessing Confusion Are Not Necessarily Suitable for Assessing Dilution

While the immediately preceding discussion refers to the control derived for assessing whether defendants' shirts caused confusion, the ProStyle court was puzzled as to why no control was used for assessing whether defendants' shirts caused a blurring of distinctiveness, a component of dilution.[85] There are two principal reasons why the assessment of blurring did not involve a control. The first has to do with the fact that different types of causal questions were being assessed. The second is that even though it was possible to create a control, a consideration of the governing legal standards reveals that such a control could not have been meaningful.

1. Causal Questions Are Not All Equivalent in Form

Suppose a glass pitcher full of water is dropped on a granite floor from a height of five feet and shatters. In order to conclude that being dropped on the floor was what caused the pitcher to shatter, do we need to set up a situation where we obtain results using a second "control" pitcher? Most people would recognize that the answer is "no." Cause and effect are obvious: Even without a control, we are able to rule out alternative explanations and conclude that being dropped (X) is what caused the pitcher to shatter (Y). Whether a second, third, fourth, etc. pitcher used for the purpose of serving as a control does or does not shatter is irrelevant to our being able to conclude that being dropped was the event that caused the first pitcher to shatter. Some causal questions are of this relatively simple form. As described in greater

85. 1999 Order, at 669.

detail below, when answering questions of the form "Does X cause Y?" controls sometimes may not be necessary.

Other causal questions—particularly those that seek to determine "what?" or "why?"—are more complicated. Compare *"Did X cause Y?"* with *"What is it about X that caused Y?"* These are different questions, and answering each requires a different type of research design. At the simplest level, the second question requires us to consider a variety of plausible alternative factors that might have been responsible. In addressing the question, "What was it about dropping the pitcher that caused it to shatter?" one could ask: Was it because of the material composition of the pitcher (being glass, it shattered; were it plastic, it would not have shattered)? Was it because of the force of its impact (if empty, the pitcher might not have had sufficient mass; being full, it had mass sufficient to generate great force at impact)? Was it because of the hardness of the surface with which it came into contact (if dropped on a plush carpet instead of a granite floor, it would not have shattered)? Was it because of some other factor(s)? Was it because of a combination of two or more of these factors? Suppose one authority contends it was simply the hardness of the surface while another argues it was the hardness of the surface combined with the force of the impact. How would one determine which of these explanations correctly identified the cause? To do so would require setting up an experiment using appropriate controls (e.g., involving dropping comparable full-to-empty pitchers from varying heights on surfaces of differing hardness).

The question of what caused the pitcher to shatter becomes even more complicated when thinking is extended beyond plausible proximal factors (such as those noted in the preceding paragraph) to distal possibilities that might have initiated a chain of causal events that culminated in the pitcher being shattered. As an example: When Tom, perhaps a bit inebriated by having too much seasonal punch, bumped into Jane, it caused some of Jane's punch to spill on the floor; when the hostess stepped on this slick spot, it caused her to slip; slipping caused her to lose her firm grip on the pitcher; losing her firm grip caused the pitcher to drop; dropping with the force that it did on the granite floor caused the pitcher to shatter. Thus, since this chain of events never would have occurred had Tom not had too much to drink, some might contend it was Tom's tipsiness that caused the pitcher to shatter.

Although additional layers of complexity can be overlaid, this example is sufficient to illustrate the ProStyle court's flawed reasoning in the present instance. As compared to, "Does X cause Y?" the question, "Does X cause Y for the reason(s) being alleged, or does X cause Y for other (non-actionable) reason(s)?" is a

considerably more complex causal question.[86] Answering the latter question will almost always require experimental designs that involve controls. In contrast, leading authorities hold that answering the former question may not necessarily require the use of controls under circumstances such as are described below.[87] In the instant matter, when it came to assessing blurring, it was the simpler ("Does X cause Y?") type of causal question that needed to be addressed. As re-phrased for the litigation context, that question was: "Does seeing defendants' garment (X) cause a blurring of distinctiveness (Y) in the minds of the relevant public?" Given a famous mark, where blurring of that mark is alleged to come from defendant's use of a combination of trademark and trade dress factors, one tests the item as a whole "as used in commerce." The test will reveal that it either does or does not cause blurring. If it does, unless other questions are at issue, there is no need to become involved in trying to parse out which of the separate elements caused the blurring.

In marked contrast to assessing whether blurring has been caused, because plaintiffs typically identify one or more specific element(s) they allege to be causing confusion, assessing likely confusion requires testing both types of questions, in seriatum. As a first hurdle, we need to address the simpler question: "Does seeing defendants' garment (X) cause confusion (Y) in the minds of the relevant public?" If the answer to this question is "yes," then we need to address the second, more complex type of causal question: "What is it about defendant's garment that causes confusion?" Phrased somewhat differently: "Does seeing defendants' garment (X) cause confusion (Y) in the minds of the relevant public *for the reasons plaintiff says it does*?" It is testing this second question that necessitates using appropriate controls.

The approach used here to assess whether ProStyle's garments caused blurring is termed a "One group post-test only" design (see earlier discussion in regard to Design 1). As the seminal thinkers in this area have written, although "*generally*

---

86. Technically speaking, answering the question, "Does X cause Y?" requires a focus on "internal validity," while answering the question, "Does X cause Y for the reasons we allege it does?" requires a focus on "construct validity." Although this important distinction is discussed in most texts on experimental research methods, the reader is directed to the seminal writings of Donald T. Campbell and colleagues. Cook et al., Quasi-Experimentation, supra n.17; Campbell & Stanley, Designs for Research, supra n.17; Thomas D. Cook & Donald T. Campbell, The Design and Conduct of True Experiments in Field Settings, in The Handbook of Industrial and Organizational Psychology 223-413 (Marvin D. Dunnette ed., 1976); Cook & Campbell, Issues for Field Settings, supra n.17.

87. The bottom line is that one can never assert that controls are always required to assess causal statements. The world of science is much more complex than that. As mentioned earlier, there is one approach, Structural Equation Modeling, that relies, quite effectively, on purely correlational (that is, non-experimental, non-control) designs for assessing mediated models of causation.

uninterpretable causally," this design *can be* effective for assessing causation under the following conditions:

> [T]he effect has to stand out, the pattern of evidence surrounding it has to be clear, the potential causes all have to be known, and auxiliary information has to be available for discriminating among alternatives when several are available.[88]

These conditions apply to the question "Did dropping the pitcher on the floor cause it to shatter?" They apply equally as well to the question "Does seeing ProStyle's garment (X) cause a blurring of distinctiveness (Y) in the minds of the relevant public?" As shown below, the obtained effect certainly stands out (being large and highly statistically significant); the pattern of evidence surrounding it is clear and compelling (also being highly significant); and the set of potential causes that apply when assessing the internal validity of a Single Group Post-Test Only Design can be identified and ruled out. The bottom line is that, notwithstanding the court's imperfect understanding of proper scientific research procedure,[89] no scientific requirement existed for controls to be used in answering the question "Did seeing ProStyle's garment (X) cause a blurring of distinctiveness (Y) in the minds of the relevant public?" Just as was the case with the glass pitcher, determining whether defendants' garments cause blurring can be accomplished without using controls.

2. Controls Are Not Always Possible Nor Meaningful

When predicated upon a false premise, what appears to be logical will often turn out to be illogical. Erroneously holding that a control designed for one specific purpose can, without

88. Cook et al., Quasi-Experimentation, supra n.17 at 517.

89. In fairness to the court, the survey report did not cite scientific treatises in support of the methodology. In fairness to this author, virtually no survey report proffered in trademark litigation does so. There are two reasons why my survey reports do not supply such citations and related technical discussion. First, my reports tend to be considerably more detailed than most, so that providing such citations and technical discussion (which, in the ProStyle matter, would have meant providing additional text equivalent to much of the present article) might be perceived as unnecessarily burdening the court. Nothing is gained from alienating a court in this way. Second, research shows that, while they support the gate-keeping role as defined by Daubert, approximately 95 percent of state judges have incorrect understandings of such fundamental concepts as "falsifiability" and "error rate," two of the four scientific concepts identified in Daubert (S. Gatowski, S.A. Dobbin, J. T. Richardson, G.P. Ginsburg, M.L. Merlino and V. Dahir, Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-Daubert World, 25 Law and Human Behavior 433-58 (2001)). Under these circumstances, there seems to be little point in providing references to technical works (e.g., Cook et al., Quasi-Experimentation, supra n.17 at 491) that most courts do not have the time or motivation to read and, if they did, likely would not understand.

consideration given to its meaning or function, automatically be used to serve a purpose for which it was not designed, the ProStyle court expressed an opinion that was both misinformed and misguided. Suppose it were to be demonstrated that a researcher-developed stimulus having one less element in common with plaintiffs' goods than does than defendants' goods will generate a certain amount of blurring. This would have no practical significance in counteracting the fact that defendant's real-world, "sold in commerce" garment (Shirt #10) does cause blurring and, hence, would be actionable.

Specifically, Shirt #7 was designed to be used as a control for assessing "net confusion" associated with defendants' Shirt #10. As used for that purpose, the causal question addressed was: "Does defendants' garment #10 cause *confusion* in the minds of the relevant public for the reasons plaintiffs say it will?" Controls are necessary for this purpose (in order to rule out alternative explanations for—or "threats" to—construct validity). In contrast, after being shown Shirt #7and then asked Question 1a ("*What, if anything, do you think of when you see this shirt?*"), the fact that some respondents answer "Green Bay Packers," though possibly useful for testing how many common elements need to be present to cause blurring, is irrelevant to answering the question: "Does Shirt #10 cause a blurring of distinctiveness?"

Failing to understand the implications of what it was doing, in its analysis of the answers to Question 1a, the ProStyle court subtracted the approximately 30 percent of the respondents shown Shirt #7 who answered "Green Bay Packers" from the more than 50 percent of respondents shown Shirt #10 who answered "Green Bay Packers."[90] Having arrived at an estimate of 20 percent "net dilution," the court commented that this 20 percent was:

> [A] figure that the court surmises would be even lower had a less misleading and more similar control than "Ellison Bay" been used.[91]

To understand why the court's comment regarding a "more similar control" is misinformed and misguided and its 20 percent figure meaningless, consider the following.

Why the court's comment regarding a "more similar control" is misguided: Earlier, a nine-point comparison chart was used to show that defendants' shirt and the control shirt were identical in terms of eight key factors and dissimilar in terms of a ninth.[92]

---

90. Recall that the study had a confusion component (Questions 2 through 4) that required using controls. Without understanding why it was improper to do so, the court relied on the data from the confusion component to the dilution component.

91. 1999 Order, at 669.

92. As noted earlier (supra n.79), the defendants' and corresponding control shirts were identical in many other respects as well.

What that illustration fails to convey is that, given nine key variables, where each is assumed to take only two levels (either "identical" or "different"),[93] more than 500 different combinations (and, hence, different potential control shirts) are possible. For purposes of illustration, let the letter "I" be used to represent "identical" and the letter "D" be used to represent "different." Applying the standard formula for calculating combinations (defined as "the number of different ways for selecting objects from a set, ignoring the order in which they are selected"[94]), there are nine different ways to create a control shirt that *differs* in only *a single* respect from defendants' shirt. These nine ways are depicted below. In terms of the nine-point comparison presented earlier, the derived control for assessing likely confusion in the first survey proffered in this matter is #8.

1	2	3	4	5	6	7	8	9
D	I	I	I	I	I	I	I	I
I	D	I	I	I	I	I	I	I
I	I	D	I	I	I	I	I	I
I	I	I	D	I	I	I	I	I
I	I	I	I	D	I	I	I	I
I	I	I	I	I	D	I	I	I
I	I	I	I	I	I	D	I	I
I	I	I	I	I	I	I	D	I
I	I	I	I	I	I	I	I	D

Given nine factors, there are many more ways—36 to be precise—in which to create a control that differs from defendants' shirt in *two* respects, yet is the same in seven respects. Given their number, only a dozen of these 36 ways are illustrated below.

93. This actually represents a gross simplification, as it fails to make allowances for the fact that, for virtually each of the nine elements, the term "Different" actually encompasses a great range of differences—from "different, but nearly identical" to "different and having nothing in common." As examples: two tee-shirts designed to look like football replica jerseys could look nearly the same (e.g., having the same number, width and spacing of stripes on the sleeves), could look nothing like each other, or could be somewhere in between; while both would be different, one shade of dark green could be highly similar to that used by defendants while another shade of dark green could differ substantially. When such variations in "Different" are factored into the equation, the universe of possible derived controls quickly goes into the millions.

94. George A. Ferguson, Statistical Analysis in Psychology and Education 71 (1959).

D	I	I	D	I	D	I	D	I	D	D	I
D	D	I	I	D	I	D	I	D	I	I	I
I	D	D	D	I	I	I	I	I	I	I	I
I	I	D	I	D	D	I	I	I	I	I	I
I	I	I	I	I	I	D	I	I	I	I	I
I	I	I	I	I	I	I	D	I	I	I	I
I	I	I	I	I	I	I	I	D	D	I	I
I	I	I	I	I	I	I	I	I	I	D	D
I	I	I	I	I	I	I	I	I	I	I	D

Additionally, there are another 84 ways in which to create a control that differs from defendants' shirt in *three* respects, 126 ways in which to create a control that differs in *four* respects, 126 ways in which to create a control that differs in *five* respects, 84 ways in which to create a control that differs in *six* respects, 36 ways in which to create a control that differs in *seven* respects and 9 ways in which to create a control that differs in *eight* respects. While there are 510 ways in which these nine key factors can be varied to create a control shirt having something in common with defendants' shirt, in terms of these nine factors, none would have been more similar to defendants' shirt than the derived control actually developed and used. Having designed a control that had eight out of nine key factors in common (amounting to a feature overlap of 8/9 = 88 percent[95]), except for using a name that changed only one letter in this ninth feature (e.g., Green Boy; Groan Bay)—an issue discussed below—it was not possible to design a control shirt having a greater number of features in common without it being identical to the infringing shirt. Thus, the ProStyle court's opinion that a "more similar control than 'Ellison Bay' [could and should have] been used" is misinformed and misguided.

Why the court's comment regarding a "more similar control" also is meaningless: Of course, it could be argued that, even though the control shirt differed in regard to only one element, that one element could have been made more similar. Instead of "Ellison" used as the control, a word more similar to "Green" in meaning (e.g., Emerald, Verdant, etc.) or sound (e.g., Grin, Groan, Groin, Great, etc.) could have been used. While, in theory, this seems reasonable, when considered in terms of what this would mean, it

95. Suppose an allegedly infringing garment contained 100 features. Further, suppose one created an exceptionally strong control garment having 99 of these 100 features in common with the allegedly infringing garment (amounting to a feature overlap of 99 percent). Now suppose we tested these garments with two randomly comprised groups and found a 50 percent confusion rate with the group tested using the allegedly infringing garment and a 49 percent confusion rate with the group tested using the control garment. Under these circumstances, clearly, it would not make sense to subtract 49 percent from 50 percent, be left with 1 percent, and then argue that the allegedly infringing garment created only 1 percent confusion.

can be shown that such a control makes absolutely no sense in terms of the controlling law.

Suppose the control shirt said "Groan Bay," "Groin Bay" (see Shirts #8 and #9) or "Green Boy," all of which sound more similar to "Green Bay" than does "Ellison Bay." Had such shirts been used in commerce, plaintiffs would have alleged that these not only were diluting via a blurring of distinctiveness, but also represented dilution via tarnishment. Suppose the control had used less pejorative similar-sounding names such as "Grin Bay," "Grain Bay" or "Great Bay." By virtue of containing virtually all the elements that, when used in combination, were alleged to dilute plaintiffs' marks, merchandise bearing these names could easily cross over the line and become the targets of legal action. Just as one cannot use another confusing mark as a control when testing likelihood of confusion test, one cannot use another diluting mark as a control when testing dilution. Since plaintiff always has the option of pursuing any third, fourth, or *n*th comer, doing so makes no sense.

Without reflecting on the implications of what it was quoting, the court also wrote:

> A control product must be a "product that is a non-infringing in this case, non-diluting product which is *similar* to the products at issue."[96]
>
> Jacoby himself has emphasized in . . . his testimony in other cases . . . that using a control group in surveys was "absolutely necessary.". . . Graham Webb Int'l v. Helene Curtis Inc., 17 F. Supp. 2d 919, 930 (D. Minn. 1998) (Jacoby criticized the opposing party's expert's survey for "failure to use third party products as a control").[97]

As earlier discussion indicated, "absolutely necessary" refers to the specific facts and circumstances surrounding that matter. Where specific elements of an item are alleged to be the factor(s) causing confusion, controls are necessary. In blurring of distinctiveness, where it is the impression conveyed by the item as a whole that is at issue, controls may not be necessary. Thus, "absolutely necessary" does not apply to all situations at all times. To reach

---

96. 1999 Order, at 670 (emphasis added).

97. Id. at 669. Elsewhere (National Football League Properties, Inc. and Green Bay Packers, Inc. v. ProStyle, Inc. and Sheri Tanner, 57 F. Supp. 2d 665, 668 (E.D. Wis. 1999)), the ProStyle court states:

> Jacoby's failure to include a control group or question in his second expert report is puzzling for several reasons. First, Jacoby himself has emphasized in both his testimony in other cases and his academic writing that using a control group in surveys is "absolutely necessary."

The court then cites the following "academic writing":

> Jacob Jacoby, Experimental Designs in Deceptive Advertising and Claim Substantiation Research, 954 PLI/Corp 167, 176 (1991).

these two words in the article cited by the court and derive from them the meaning extracted by the court one would have to ignore the clearly stated qualifying language in the opening sentence of that very same section that states: "Though there are special instances where *testing causal questions may* not *require the use of control groups*, their use is *generally* called for" (emphasis supplied). One only arrives at the "absolutely necessary under all circumstances" meaning if, for whatever reason, one chooses to ignore the clearly-stated qualification.

Further, in the present situation, no "similar" "third party" products existed in the marketplace. Had such products been available, it is likely that they, too, would have become the focus of a lawsuit. True, one could develop a derived control (e.g., "Groan Bay") that would, because of having all the other features in common, be maximally similar to both plaintiffs' and defendants' garments. But, what would it mean if such a derived, non-real-world control evoked in the minds of the relevant public a considerable number of associations to plaintiff? It would simply mean that, if one tried, one could create a fictitious product that would be diluting. The only thing that would prevent it from being judged such would be the fact that it failed to satisfy that portion of Section 43(c) that requires diluting products to involve "another's *use in commerce* of a mark or trade name." There can be little doubt that, had such shirts appeared in commerce, plaintiffs would have filed suit and likely prevailed.

For this reason, the ProStyle court's argument necessarily fails. Just as one cannot use a third party's confusing product as a control when testing whether a second party's product is confusing, one cannot use a third party's diluting product as a control when testing whether a second party's product is diluting. Suffice it to note that because it is possible to subtract the 30 percent association with a derived control having no presence in the real world from the more than 50 percent association with an item being sold in the real world and arrive at 20 percent does not invest this latter percentage with any meaning. It reflects only the manipulation of numbers without any consideration of what these numbers mean.

F. Controls and the Necessity for Comparisons

In support of its position that "controls" were required, but without understanding its full import, the ProStyle court cited the following from the FJC's Reference Manual on Scientific Evidence as support.

> David H. Kaye and David A. Freedman Reference Guide on Statistics, in Reference Manual on Scientific Evidence. At 348-49 (1994) ("Outcome figures from a treatment group without a

> control group reveal very little and can be misleading. Comparisons are essential.")[98]

Even if one failed to understand that "can be" was a phrase deliberately chosen to convey a precise meaning, most would understand that "can be misleading" does not mean the same thing as "necessarily is misleading," the "spin" given this phrase by the ProStyle court.

More importantly, the court grasped the basic concept—namely, that, upon seeing defendants' garments, the fact that more than 50 percent of the respondents thought of plaintiffs' needs to be compared. But compared to what?[99] What the court apparently failed to grasp was that several types of comparison not only are possible, but meaningful. At least three broad approaches can be identified.

Compare findings obtained using the allegedly infringing item to findings obtained using a control. In the best of all worlds, one would compare the findings obtained with the treatment to the findings obtained with a control. Although this is what the court called for, as indicated, the control must be appropriate and make sense. If not, it is nothing but a false control—something that, suggesting rigor and meaning when neither exists, likely would produce misleading results.

In the present instance, there were no natural, third-party, real-world controls—at least none similar enough to defendants' garments that they could be expected to elicit any association with plaintiffs. Had such third-party garments been used as controls, being weak controls, defendants would have argued that they constituted a trick designed to obtain low-to-non-existent levels of confusion with the control which, when subtracted from the test of defendant's garment, would have produced inflated estimates of "net" blurring.

When no natural controls are available, one begins to consider developing a derived control (recognizing that, as with a natural control, it must not be bogus but must make sense). However, as the number of elements possessed in common by plaintiff's (first comer's) and defendant's goods increase, the researcher is always faced with a problem. As the number of common elements (feature overlap) between the control and the first comer's mark or dress increases, the control itself can be expected to create a certain amount of confusion. Thus, how many elements can or should be incorporated in a derived control before it loses its value as a control? Regardless of the answer, this question may well be meaningless. The fact that a control shirt having considerable

---

98. 1999 Order, at 668-69.

99. To understand this point, imagine a one-armed fisherman extending his one arm to indicate that he "caught a fish this long."

feature overlap can be created and may cause as much as or more blurring than defendants' goods cannot reasonably have any bearing on whether action can be taken against defendants' goods. If and when goods such as the derived control are "used in commerce," plaintiff would be fully justified in pursuing those who developed and used such marks.

In sum, in the present instance, there was no basis for drawing comparisons to a meaningful control. Comparison to a natural control would have yielded "net" results strongly in favor of plaintiff (and thus risked being labeled "a trick"), while reliance upon a derived control that possessed many of the infringing elements would be irrelevant.

Compare associations evoked by defendant's mark(s) or dress with those evoked by plaintiff's authentic famous mark(s) or dress. A second meaningful comparison for a test of blurring would be to compare the results of testing defendant's goods not to a control, per se, but to the associations evoked by plaintiff's authentic famous mark(s). Famous marks and logos—as examples, ROLLS ROYCE and the NIKE swoosh—are not famous when first introduced, but become famous over time. A famous mark is thus one that already has been tested by time, so that seeing or hearing it now evokes thoughts of a singular source in the minds of a large proportion of the relevant public.[100] By definition, that is what makes it famous. In testing whether an item causes a blurring of distinctiveness, the essential question is whether defendant's marks cause thoughts of plaintiff (or its goods) to be evoked in the minds of a substantial number of the relevant public—regardless of whether or not some third party's goods do so as well. In this sense, plaintiff's famous mark serves as the comparison and the extent to which defendant's goods cause blurring (namely, thoughts of the plaintiff to be evoked) can be compared to the original. It matters not whether other marks or dress may yield similar findings; if ever used in commerce, they could be actionable as well.

As appropriate, employ the comparative logic used in quasi-experimentation. A third meaningful basis for comparison involves employing the logic and approach applied with quasi-experimental designs. As the leading authorities on this subject have stated: "The major alternative to control via design is control via measurement and statistical adjustment."[101] When considering the latter, "causal inference is strengthened by increasing the number,

100. According to the House Report accompanying the Dilution Act, a claim for dilution "applies when the unauthorized use of a famous mark reduces the public's perception that the mark signifies something unique, singular or particular." See HR Rep. No. 104-374 at 3.

101. Cook et al., Quasi-Experimentation, supra n.17 at 570. Note that the phrase "control by design" refers to designing an experiment to incorporate control groups or control conditions.

complexity, and specificity of the data-based predictions derived from a causal hypothesis. . . ."[102] Accordingly, comparing the obtained findings to what was predicted, what specific "data-based predictions derived from a causal hypothesis" can be assessed by the data adduced in the present situation?

In arguing that ProStyle's garments caused a blurring of distinctiveness, plaintiff relied on data derived from a single question, *"What, if anything, do you think of when you see this shirt?"* As this is a completely open-ended question, respondents could have given any sort of answer, including answers that had nothing whatever to do with football, professional football, the NFL, or the Green Bay Packers. Thus, one meaningful comparison is to compare what plaintiff predicted (namely, exposed to defendant's shirts, consumers would be caused to think of and draw association to the Green Bay Packers) to the other answers given by the respondents. Note that the fact that this question is completely open-ended enables us to rule out the rival explanation that it was the question itself that caused respondents to give the answers they gave. Also, as it was the very first question asked, we can rule out the rival explanation that the answers to this question might have been influenced by the respondent having been asked and answered prior questions.

Given an open-ended question used to assess the causal hypothesis "Exposure to defendant's allegedly diluting shirts will cause consumers to think of, and draw association to, the Green Bay Packers," at least three specific data-based predictions can be derived:

1. More associations will be made to the Green Bay Packers than to anything else.[103] (If this hypothesis is refuted, the basis for a dilution claim likely would be eviscerated.)

2. Given that associations may also be made to other NFL teams, a significantly higher percentage of the associations would be made to the Green Bay Packers than to any other NFL team. (If this hypothesis is refuted, this would also seem to quash any basis for a dilution claim.)

---

102. Id. at 571.

103. In scientific parlance, the situation can be described as follows. The independent variable (a presumed cause consisting of exposure to defendants' shirt coupled with being asked, *"What, if anything, do you think of when you see this shirt?"*) is completely under the control of the investigator. There is an a priori prediction that exposing respondents who are members of the relevant universe to this independent variable will produce a specific and singular effect (namely, causing these respondents to draw associations to the Green Bay Packers and nothing else). Although an almost infinite number of other answers are possible, any such answers would represent non-predicted effects against which the predicted effect can be compared. Note that the a priori nature of the hypothesis precludes the post hoc sifting through the data for the purpose of verifying other previously unconsidered causes.

3. Regardless of where tested, when exposed to defendant's garments, respondents at all testing sites outside of Wisconsin would say defendant's shirts made them think of the Green Bay Packers to a significantly greater extent than they would say it made them think of their local NFL team. (If this hypothesis is refuted, this would also appear to quash any basis for a dilution claim.)

What do the data reveal regarding these specific predictions?

When shown defendants' shirt (the presumed cause of blurring) and asked, *"What, if anything, do you think of when you see this shirt?"* respondents gave a variety of answers, with many giving more than one. A number of answers were quite general and had nothing whatever to do with sports.[104] A few respondents gave sports-related answers not tied in any direct way to football.[105] Yet others mentioned football either in general (e.g., football; football season; football team), or without indicating a level of play (high school, college, non-professional, professional) or a specific team. Tellingly, none of the answers described above mentioned the names of any entities whatever.

Other respondents said that defendants' shirt made them think of the National Football League, a named entity and one of the plaintiffs in this matter. Although these latter answers might be taken as evidence of blurring, in the interest of being conservative, these findings were ignored. Consider, now, the three hypotheses articulated above.

A table summarizing the data from the dilution report proffered in the ProStyle matter is provided below.[106] While separate analyses can be provided for each of the nine percentages, in the interest of simplifying discussion, focus on the bottom-right figure of 52 percent. This percentage is the overall average for respondents tested under the point-of-sale protocol and others tested under the post-sale protocol, and for those tested on the first of defendant's shirts and others tested on the second of defendant's shirts.

---

104. As examples: It's a tee-shirt; It's big; Nice looking shirt; Plain looking shirt. Undershirt; My husband; Looks comfortable; Spending a Sunday watching the game and getting together with friends; Having a party; Night shirt; Fairly heavy; I like the style; Emblem; Don't really care for the color; Looks nice; The size; Looks comfortable; Short sleeve or long sleeve; The beach; Warmth; The logo; Cold weather; Warm; That it's all cotton; Something to lounge around in; I like mine extra large; A bay; I don't like the collar on it; It's a sporty, nice looking shirt.

105. As examples: Sports; Sports insignia; I don't like this style of shirt or the team on it; Team spirit.

106. These percentages have been calculated using all 324 respondents as the base, that is, without disregarding any of respondents giving irrelevant responses such as those discussed above.

Percent of Respondents Saying Defendant's Shirt Made Them Think of the Green Bay Packers[107]

	Point-of-Sale	*Post-Sale*	*Average of PoS + PS*
	%	%	%
Shirt Type 1	62	36	52
Shirt Type 2	56	44	51
Shirt Types 1 and 2	59	40	52

Consider Hypothesis 1: "More associations will be made to the Green Bay Packers than to anything else." When 52 percent is compared to 0 percent (the latter being the percent of respondents who answered giving the name of a non-NFL related named entity), from a statistical standpoint, the evidence is exceptionally significant and strongly confirms Hypothesis 1. The likelihood of obtaining such a result by chance is less than one out of 1,000.

Consider Hypothesis 2: "Given that associations may also be made to other NFL teams, a significantly higher percentage of the associations would be to the Green Bay Packers than to any other NFL team." The National Football League consisted of 30 member teams dispersed across the continental United States. If respondents had answered Q1a by naming these 30 teams at random, we would predict each team to be mentioned approximately one out of 30 times, or 3.3 percent of the time. Yet *not a single one of the 324 respondents* tested for blurring mentioned *any* of the NFL's 29 other teams. In contrast, 168 (52 percent) of these 324 respondents explicitly mentioned the Green Bay Packers.[108]

As the leading thinkers on experimental design state,[109] the "one group, post-test only" design is capable of yielding valid causal inferences when (1) the effect stands out, (2) the surrounding pattern of data is clear, and (3) the other potential

---

107. Jacob Jacoby, "The Fame of the Green Bay Packers as Embodied in Its Registered Marks, and the Extent to Which ProStyle Merchandise Dilutes and Is Likely to Cause Confusion with These Marks." September 1998, at 27. Note that the nine percentages presented in this table become appreciably higher when evidence regarding blurring is not confined to Question 1a (as is the case for the data provided above), but gathered from the answers to all questions the respondent was asked.

108. As examples: Packers; Packer football; Green Bay Packers monogram; Someone who is a Packers fan; The color, Packer green; Green Bay Packer logo and football fans would like it; My dad because he likes the Green Bay Packers; It's a Packer shirt.

109. "[T]he effect has to stand out, the pattern of evidence surrounding it has to be clear, the potential causes all have to be known, and auxiliary information has to be available for discriminating among alternatives when several are available." Cook et al., Quasi-Experimentation, supra n.17 at 517.

causes are known and can be ruled out.[110] In terms of the first of these criteria, not only is 52 percent "large" but, when compared either to 0 percent (the percent of times these respondents were observed to have mentioned the names of any other entities or teams in any sport at any level and at any place elsewhere in the country), or 3.3 percent (predicted if we use the "other NFL teams at random" assumption), 52 percent "stands out." To dot the "i," when tested for the significance of the difference, 52 percent turns out to be statistically different from 3.3 percent[111] at well beyond the $p < .0001$ level, a less than one in a thousand chance.[112] In other words, Hypothesis 2 is confirmed.

Consider Hypothesis 3. "Regardless of where tested, when exposed to defendant's garments, to a significantly greater extent, respondents at all testing sites outside of Wisconsin would say defendant's shirts made them think of the Green Bay Packers, not their local NFL team." Although two-thirds of the respondents came from Wisconsin, the other one-third—more than 100 respondents—came from Chicago (the home city of the NFL's Chicago Bears) and Minneapolis (the home city of the NFL's Minnesota Vikings). Examination of the data revealed that, upon seeing defendants' shirts, not a single one of the respondents in Chicago or Minneapolis were caused to think of either the Chicago Bears or Minneapolis Vikings. Instead, approximately 50 percent of the respondents at each site said that defendants' shirt made them think of the Green Bay Packers.

With regard to the second criterion—"the surrounding pattern of data is clear"—the data provide a compelling picture. When analyzed via the standard Chi Square statistic used to determine

---

110. Operating with its superficial and flawed understanding of the concept of controls, oblivious to the acceptance of this approach by the leading thinkers in the relevant scientific community, the ProStyle court commented:

> Plaintiffs next argue that "the open-ended dilution question itself contains an infinite number of controls—in the very answers provided by respondents."... The court doubts whether a single-question survey could contain its own controls when the whole concept of the "control" is that there must be separation, e.g., into groups or between questions, and then comparison between the separated parts.

57 F. Supp. 2d at 670. Notwithstanding the court's doubts, one can and often does separate the answers given to questions and then makes important comparisons between the separate parts. To see that this is so, consider the typical likelihood of confusion survey where answers to the question "Why do you say that?" are separated and only those that reflect confusion due to the alleged cause are counted, while all other answers are not.

111. The original report provided data for 648 respondents, including 324 control group respondents used for assessing likely confusion. Not being relevant for assessing blurring, the data for these 324 respondents were removed from the second report. This left the 324 respondents who were tested to determine whether defendants' shirts caused a blurring of distinctiveness.

112. As compared to the level of significance used in most scientific research, 1 out of 20 (known as the ".05 level of significance"), 1 out of 10,000 can be seen to be "off the charts." Clearly, the effect "stands out." The "pattern of surrounding evidence" consideration is discussed below.

the significance of the difference between an expected pattern of 3.3 percent mentions for each NFL member team and an empirically observed pattern (which, in this case, consists of twenty-nine names at 0 percent vs. one name at 52 percent), we obtain a Chi Square value that is so highly significant that it is likely to occur by chance at less than one in ten thousand. In other words, the pattern of obtained findings could not be any more clear, compelling and convincing.

Also recognize that, for ease of discussion, attention has been focused on only one of the nine percentages reported in the above table. The same three hypotheses can be tested for each of the other eight percentages. When this is done, in each and every instance, the hypotheses are confirmed at the same highly significant levels. Thus, in all, there were (9 x 3 =) 27 tests of specific hypotheses. In any one of these tests, the hypothesis could have been disconfirmed if the findings were found likely to occur less than one out of twenty times. Yet not once were any of the hypotheses disconfirmed by the data. The specific hypotheses were confirmed in all twenty-seven instances, again, a remarkable non-chance result that is significant at an extraordinarily high level.

Finally, what about the ability to identify and rule out other potential causes that might be used to explain the findings? When one evaluates all the "threats to internal validity" (i.e., potential alternative explanations for why the data suggest that X caused Y) that can apply to a "One Group Post-Test Only Design," it can be seen that none of these alternatives could cause the obtained effect.[113] Hence, this leaves the plaintiff's explanation—that seeing defendants' garments causes a blurring of distinctiveness—as the most reasonable remaining inference.

The data thus reflect precisely the kind of situation deemed acceptable when using a "One Group Post-Test Only Design" for assessing causation.

> [T]he effect has to stand out [at 52 percent, it certainly does], the pattern of evidence surrounding it has to be clear [at 0 percent for each of the other NFL teams, the pattern is crystal clear], the potential causes all have to be known [all other threats to internal validity have been ruled out], and auxiliary information has to be available for discriminating

113. These alternative threats to internal validity are: Maturation (changes in the respondents between the time they were exposed to defendants' garments and at the time when their associations were measured); Testing (being asked the question a number of times, the respondents have taken the opportunity to look up the answer after one of these occasions, but before answering the question on another occasion); Instrumentation (effects due to the measuring instrument); or Ambiguity about the direction of causal inference (here, there is no reason to believe that the respondents were thinking about the Green Bay Packers before being shown defendants' garments and asked to indicate what they thought). See Cook et al., Quasi-Experimentation, supra n.17 at 500-01.

among alternatives when several are available [a criterion not applicable in the present instance].

Not understanding the clear and compelling nature of these data, the court commented:

> [T]he court finds it utterly unremarkable that more than half of all Wisconsinites polled shortly before the Packers' first Super Bowl appearance in nearly 30 years "thought of" the Packers when shown green and gold shirts that said "Green Bay Football" or "Green Bay P.". . . The court surmises that, in this state and at such a time of Packer-related frenzy, the questioners could have shown the survey respondents a green and gold Blarney stone (an example of a non-diluting use) and more than half of them would have thought of the Packers.[114]

By divorcing the colors from their context, it seems the ProStyle court misrepresents the central *legal* issue. Plaintiffs never did, nor ever would, have argued for protection of the green and gold color combination in isolation. Rather, plaintiffs argued for protection when this combination of colors was used for goods being offered for sale "in commerce," and these goods *also* carried the name Green Bay, a 15-inch letter "P" adjacent to Green Bay, the phrase "Go Pack Go," some football indicia, etc. If the green and gold Blarney stone had also said Green Bay, contained a large letter "P" and/or football indicia and satisfied the Lanham Act's Section 43(c) requirement of being "another person's commercial use in commerce of a mark or trade name," it is not at all clear that the ProStyle court's Blarney stone would constitute "an example of a non-diluting use."

Regardless, whether or not said Blarney stone was found to be diluting, it is far from clear that this would be relevant. The fact that shirts can be envisioned and devised that may cause as much or more dilution than defendant's marks should have no bearing on whether action can be taken against defendant's real-world usage. If and when such devised marks are "used in commerce," according to the dilution statute, plaintiff would be fully justified in pursuing the users of such marks as well. Thus, it seems that about the only thing the court's Blarney stone example does is attest to the very high level of fame attaching to plaintiff's distinctive trade dress and to goods adorned with this dress.

H. Controls for Testing Advertising Copy and Package Labeling

The typical trademark matter focuses on a single name, symbol or limited phrase (such as a slogan). A more complicated situation may arise when the matter at issue concerns the

114. 1999 Order, at 669.

(generally textual, but occasionally graphic) content of advertising or package labeling, and the meanings conveyed by this content. The types of controls generally used when testing whether or not advertising is deceptive include the following.

(1) Use Design 5, as discussed earlier. That is, while those in the Experimental group are exposed to the contested ad (brochure, periodical, package, etc.), those in the Control group are not exposed to any ad (brochure, periodical, package, etc.), and the "effects" are assessed for both groups at approximately the same time. Note that this "post-only" design is capable of accommodating the issue of "pre-existing beliefs."[115] Inasmuch as subjects randomly assigned to the Control group could be assumed to have, on average, the same "pre-existing beliefs" as those assigned to the Experimental group, this will "net" out when the findings from the Control group are subtracted from those for the Experimental group. A problem sometimes created when relying on a "no stimulus" control is that the questions asked of respondents may have to be worded a bit differently for those in the Experimental vs. those in the Control group, thereby making the experiences across the two groups at the point of measurement less "standardized."

(2) Use Design 5 as discussed earlier, except instead of not exposing Control group respondents to an ad (brochure, periodical, package, etc.), do expose the Control group respondents to some ad (brochure, periodical, package, etc.). Depending upon the claims at issue, the control communication may take any one of several forms. At least two types of natural controls often are possible. In many instances, it is preferable to have the control communication come from the same source and for the same product and brand, but be a version that does not contain the allegedly deceptive or confusing verbiage. Another less frequent (and generally less acceptable) alternative is to have the control be a comparable ad, package, etc. for the same product, but from a competing source.

Insofar as print media are concerned, a great number of possibilities exist with regard to developing derived controls. At one end of the continuum, a derived control might involve making minimal changes. For example, if a name, word or brief phrase is at issue, the same ad might be used, but with a different term substituted for the allegedly deceptive or confusing term. At the other extreme, one might create a control ad from scratch. Two in-between options are using a "tombstone ad," essentially an ad containing nothing other than the name of the source and possibly the brand name of the product, or a "purged" control. A purged control is one which uses the same ad, except that the verbal (or

115. Cases discussing "pre-existing beliefs" include American Home Products v. Procter & Gamble Co., 871 F. Supp. 739 (D.N.J. 1994); Johnson & Johnson-Merck Consumer Pharmaceuticals Co. v. SmithKline Beecham Corp., 960 F.2d 294, 301 (2d Cir. 1992).

sometimes graphic) content at issue is removed and nothing else inserted to take its place.

Although this writer has advocated "purged controls,"[116] purged controls have limited applicability. While it may be effective in those instances where just one or a few words are at issue, it becomes problematic when more textual material, and especially when more than one concept, is involved. For example, in the FTC v. Kraft matter described earlier, the FTC alleged that approximately 30 percent of the verbal content of the Kraft communications was in some way responsible for creating the alleged deception. Hence, simply removing the word "calcium" from these ads would have not been sufficient.

As purged controls are sometimes used when assessing likelihood of confusion, brief discussion of the problem is merited. Suppose 30 percent of the verbiage (or 30 percent of the trade dress) contained three distinct components—labeled A, B, and C—and it was unclear just which one, *or combination*, of these components was responsible for causing the alleged deception. Given three factors, definitively addressing the question of causality would require a "factorial" experiment involving seven separate test groups (each exposed to a different combination of these message components: A+B+C; A+B; A+C; B+C; only A; only B; only C). Also useful would be an eighth group exposed to a version that contained none of these potentially misleading components. Not many advertisers (or trademark owners) would be willing or able to fund such research.

Last, recognize that, because they are dynamic media, constructing derived controls for use with allegedly deceptive broadcast (TV or radio) or dynamic Internet advertising generally raises many more practical and conceptual difficulties than is the case with print advertising.

(3) Rely on Internal Controls. When neither natural nor derived external controls are practical, an alternative is to rely on internal controls. For an illustration, the reader is directed to the earlier discussion of Gillette v. Wilkinson Sword (see supra Part IV.D.).

VII. CONCLUSION

This article has but scratched the surface of experimental designs as data-gathering strategies for assessing causal propositions. The topic is both detailed and complex. Many important issues have not been touched upon. However, an

116. Jacob Jacoby, Defining and Measuring Misleading Advertising, Final report to the Division of Drug Advertising, Food and Drug Administration (1974). For a condensed version, see Jacob Jacoby and Constance B. Small, The FDA Approach to Defining Misleading Advertising, 39 Journal of Marketing 65-68 (1975).

introduction has been provided to the vocabulary, logic and basic forms of experimental design, and case law has been described to illustrate how these have been and can be used to assess Lanham Act issues pertaining to causation. Additionally, it is hoped the reader will come away appreciating the following wisdom.

First, although fully experimental designs are often the preferred strategy, not all real-world questions are amenable to the application of such designs. Scientifically acceptable alternatives do exist, including quasi-experimental designs (some of which do not rely on control groups) and non-experimental Structural Equation Modeling that generally does not involve control groups. Second, especially for the non-scientist, the fundamental fully-experimental designs can sometimes appear so complicated that, as witnessed by the Seventh Circuit's discussion of "Horses" v. "Leopards" (see Part VI.A., above), even the best and the brightest may experience difficulty parsing out their significance and meaning. Third, with a domain as complex as experimental design, it would be naïve for anyone—counsel or court—to think that everything that needed to be said was encompassed in a few pages of the Federal Judicial Center's Reference Manual on Scientific Evidence (or even in this article, for that matter). Last, the world is complex so that there are no easy answers. If there were, there would be no need for the application of rigorous scientific procedures and scientists schooled in their application and interpretation.

The interpretation of data derived from the use of experimental designs can be equally as challenging. Consider the following tale. There once was a man who, after drinking gin and tonic at a friend's home, became intoxicated and suffered a serious hangover the next day. Sitting in a tavern a few days later he decided to try bourbon and tonic, again becoming intoxicated and suffering a serious hangover the next day. A few days later, in the privacy of his own home, he tried vodka and tonic. When, once again, he became intoxicated and suffered a serious hangover, he concluded that since the only constant across all three situations was tonic, tonic causes drunken stupors and hangovers! The implication of this tale: the design and interpretation of experimental studies is sufficiently complicated not to be left to those inexperienced in these matters.
