

## A Practical Guide to Experimental Advertising Research

Patrick T. Vargas, Brittany R. L. Duff & Ronald J. Faber

To cite this article: Patrick T. Vargas, Brittany R. L. Duff & Ronald J. Faber (2017) A Practical Guide to Experimental Advertising Research, Journal of Advertising, 46:1, 101-114, DOI: [10.1080/00913367.2017.1281779](https://doi.org/10.1080/00913367.2017.1281779)

To link to this article: <http://dx.doi.org/10.1080/00913367.2017.1281779>



Published online: 22 Feb 2017.



[Submit your article to this journal](#)



Article views: 748



[View related articles](#)



[View Crossmark data](#)

# A Practical Guide to Experimental Advertising Research

**Patrick T. Vargas and Brittany R. L. Duff** 

*University of Illinois at Urbana–Champaign, Urbana, Illinois, USA*

**Ronald J. Faber**

*University of Minnesota, Twin Cities, Minneapolis, Minnesota, USA*

---

**Experiments are conducted to help establish cause-and-effect relationships, and they can be powerful tools for doing so. We review fundamental concepts for conducting experimental advertising research. Good experimental research involves careful consideration of independent and dependent variables, and what they are supposed to represent. To this end, we review threats to construct validity as well as offer some suggestions on how to think about external and ecological validity in ad research. We review three quasi-experimental research designs and three simple, randomized experimental designs, along with more complex factorial design experiments. Finally, we discuss ethical considerations and the crucial role researchers play in maintaining research integrity.**

---

Experiments are designed and conducted to test whether, and to what extent, one thing causes another. Advertising researchers have used experiments to examine whether point-of-purchase advertising influences sales (e.g., Caballero and Solomon 1984; Greco and Swayne 1992), whether mood induced by a TV program affects processing of subsequent advertising (Cline and Kellaris 2007; LaTour and LaTour 2009; Shapiro and MacInnis 2002), whether repetition is an effective way to improve memory for brands (Nordhielm 2002), and whether increasing the salience of consumers' ethnic identities affects responses to ads with patriotic appeals (Yoo and Lee 2016). Causal relationships are implicit in each of these relationships, and the best tool

researchers have for determining causal relationships is experimental research.

An experiment involves an investigator manipulating and controlling one or more potentially causal variables (independent variables) and then observing the corresponding differences in the outcome: the dependent variable(s). By controlling the situation, the researcher can eliminate some of the external conditions that might confound or confuse the results. For this reason, experiments are typically the preferred research method for demonstrating causation. For example, if a researcher wanted to know whether point-of-purchase ads make people more likely to buy a product, one option would be to conduct a survey asking people whether they are more likely to buy products based on a point-of-purchase ad. People would very likely provide sensible responses, but there is a large body of research showing that people are not very good at knowing the true reasons why they do the things they do (e.g., Nisbett and Wilson 1977). Imagine a researcher asking a woman why she uses the brand of laundry detergent she does. She is likely to say because it cleans her clothes well. But is this really the answer? As likely as not, the real reason is because this is the brand her family used growing up, or it's the brand her friend or roommate used and she found it was good enough. In fact, she may have never really thought about it until she was asked, and then she just wanted to come up with a reasonable-sounding answer. Now think about a very different question: Imagine asking a heterosexual consumer how he feels when he passes a gay or lesbian couple engaged in a public display of affection or sees an advertisement featuring a gay or lesbian couple (Bhat, Leigh, and Wardlow 1998). Few people like to think of themselves as prejudiced, and even if they know that they are prejudiced they may not want to admit it. By conducting an experiment, researchers can avoid relying on potentially faulty memory, having people give socially desirable answers, or having people come up with some answer even when they do not know why they did something. Experimental research helps address the problems inherent to asking people "why" questions, the possibility of alternate explanations, and the problem of what's really causing what.

---

Address correspondence to Patrick T. Vargas, University of Illinois at Urbana–Champaign, Department of Advertising, 810 S. Wright St., MC-462, Urbana, IL 61801. E-mail: patrick.vargas@gmail.com

Patrick T. Vargas (PhD, The Ohio State University) is a professor of advertising, College of Media, University of Illinois at Urbana–Champaign.

Brittany R. L. Duff (PhD, University of Minnesota, Twin Cities) is an associate professor of advertising, College of Media, University of Illinois at Urbana–Champaign.

Ronald J. Faber, (PhD, University of Wisconsin, Madison) is professor emeritus of mass communication, School of Journalism and Mass Communication, University of Minnesota, Twin Cities.

Experimental research also helps test and refine theory. A theory is a group of ideas used to explain events and make specific predictions about future events. Scientific theories are testable and must be stated in a way that makes them falsifiable (Popper 1959). A theory should allow researchers to make specific predictions, hypotheses, about what will occur under precise conditions. A theory that is not falsifiable, that cannot be demonstrated to be incorrect, leaves researchers trapped within that set of beliefs and prevents progress. For example, if someone believes that advertising always causes more favorable attitudes toward advertised brands, that person would be inclined to rationalize or explain away any situation that was inconsistent with that belief (e.g., if advertising did not increase favorable attitudes it must be because the viewers were too stupid to understand the message) to maintain the original theory. Experiments are one set of tools that researchers can use to test hypotheses derived from theories to determine whether they hold up to scrutiny (and should therefore be retained) or fail (and should be modified or discarded).

A particularly good use of experimental research is testing theories that make competing predictions (Platt 1964). Consider, for example, trying to understand how mood induced by a happy or sad TV program affects consumer responses to happy and sad ads (Kamins, Marks, and Skinner 1991). According to mood congruency theory, when people are happy they like everything more, and when they are sad, everything seems less good. Because people are happy when watching a happy TV program, all ads (even sad ones) are best placed in happy programs. However, mood *consistency* theory predicts that people will prefer an ad that is consistent with, and maintains, the mood they are already in. Thus, if one needs to place a sad ad (e.g., a plea to donate money to help starving children) congruency theory would predict it will do better placed in a happy program, while consistency theory would predict it will do better placed in a sad program. Kamins, Marks, and Skinner (1991) tested these opposite predictions in an experiment and found stronger support for consistency theory. Experiments do not always definitively resolve competition among theories (see Greenwald 2012), but they do help verify the acceptability of theories.

## KNOWLEDGE AND CAUSAL INFERENCE

What, exactly, are causes and effects? Generally speaking, causes are things that produce (e.g., ideas, advertisements) and effects are the things produced (e.g., attitudes, beliefs, intentions). According to John Locke, “A cause is that which makes any other thing, either simple idea, substance, or mode, begin to be; and an effect is that, which had its beginning from some other thing” (Locke 1689, cited in Shadish, Cook, and Campbell 2002, p. 4). Establishing causal relationships requires meeting three criteria (Mill 1843). First, the cause and the effect must vary together (i.e., be correlated). Take the well-known example of advertising budgets being correlated

with sales. While these variables are related, it does not necessarily mean that larger advertising budgets cause more sales. Brands often set their advertising budgets based on a percentage of last year’s, or projected future, sales. Alternatively, both budgets and sales might be related to some third thing, such as a booming economy. Thus, even if two variables are correlated there is no guarantee that they are causally related.

Second, the cause must precede the effect. This is known as temporal precedence, and might seem fairly straightforward, but it is not always so. Imagine a study where people are shown lots of ads. Afterward, they are asked which ads they recall and then asked about their brand attitude for each brand advertised. Researchers find that the ads that were recalled by more people have higher brand attitude scores. Can the researchers say that ad recall led to (i.e., caused) higher brand attitudes? The answer is no, they cannot, because it is also possible that the brands people like more (i.e., brands given higher attitude scores) are more likely to have been noticed and remembered. Thus, while the two variables are correlated, establishing causation would require an experiment that explicitly manipulates, or controls, recall while assessing effects on brand attitudes.

This leads to the third requirement in establishing causal relationships: attempting to rule out plausible alternative explanations for the observed relationship. This is known as the internal validity criterion. An experiment with high internal validity—it controls for, or rules out, alternate causes—is one for which researchers can be relatively certain about the cause-and-effect relationship. Without internal validity, researchers cannot determine whether an outcome was produced by one factor or another, so it is vitally important to understand.

According to Mill (1843), a logical basis for the justification of claimed causal relationships between variables can be established with necessary and sufficient causes. This is accomplished with three straightforward ideas: the methods of agreement, difference, and concomitant variation. First is the method of agreement: “If  $X$ , then  $Y$ .” If  $Y$  occurs where  $X$  is present each time, then it can be said that  $X$  is a sufficient cause of  $Y$ ;  $X$  is adequate for bringing about the effect. The second idea is the method of difference: “If not- $X$ , then not- $Y$ .” If  $Y$  does not occur when  $X$  is absent, then  $X$  is a necessary, or essential, condition for causing  $Y$ . Third, is the method of concomitant variation: “Variations in  $Y$  are functionally related to variations in  $X$ .” When conducting experiments, researchers employ Mill’s method of establishing necessary and sufficient causes via experimental and control groups.

A good example to illustrate Mill’s logic comes from research a century ago on pellagra, a disease that killed approximately 100,000 people in the early 1900s. This disease was common among poor people in the American South who lived with inferior plumbing and sewage disposal but was far less common among wealthier people with superior plumbing and sewer systems. Doctors studying pellagra were confident

that some microorganism caused the disease. However, Joseph Goldberger believed otherwise; he thought it was caused by the high-carbohydrate, low-protein diet on which poor people subsisted. Goldberger bravely tested the microorganism hypothesis by having himself and some assistants inject and ingest the blood and secretions from pellagra patients. Neither Goldberger nor his assistants became ill, thereby ruling out the sewage-related microorganism account of the disease. Next, to test his diet-based account, Goldberger asked a volunteer group of prisoners (who had access to adequate sewage systems) to subsist on a high-carb, low-protein diet, and another volunteer group to subsist on a more balanced diet. Within five months the high-carb, low-protein diet prisoners were suffering from pellagra, while the balanced diet prisoners remained healthy (Stanovich 2010). Thus, Goldberger saw the correlation of poverty and pellagra (method of agreement) but sought to rule out alternative explanations by testing the idea that pellagra was infectious and passed via poor sanitation (method of difference: exposure to microorganisms did not cause pellagra), and then manipulated another potential cause, diet, while keeping sanitation constant (concomitant variation). By carefully ruling out microorganisms and establishing diet as the cause of pellagra Goldberger helped protect his conclusions during the ensuing outcries and attacks by leaders of states with high pellagra rates who worried that the publicity from his findings would hurt their states' images (Elmore and Feinstein 1994).

### INDEPENDENT AND DEPENDENT VARIABLES

Experiments allow researchers to control and manipulate presumed causal factors or, as they are known in experimental research, independent variables (IVs). Researchers observe how IVs affect outcomes or dependent variables (DVs). IVs and DVs are referred to as variables because they are subject to variation. In advertising research, typical IVs may include things like mood (happy versus sad; Zhao, Muehling, and Kareklas 2014); involvement level (Putrevu and Lord 1994); the effects of particular advertising elements, such as a celebrity spokesperson's attractiveness (Kamins 1990), the amount of white space in an ad (Pracejus, Olsen, and O'Guinn 2006), or the type of metaphor in an ad (Chang and Yen 2013). Common advertising DVs include outcome variables such as ad attention, memory, attitude, liking, and actual or intended behavior.

Generally, in advertising research IVs are manipulated or controlled to be discrete variables while DVs are continuous, although they do not have to be this way. One reason experimental IVs are typically discrete is because using a continuous IV would require as many experimental groups as there are levels of that particular variable. When continuous IVs are used in advertising research, they are typically assessing something like a personality trait or attitude. For example, a researcher conducting an experiment on the effects of mood on memory for an ad might randomly assign people to either a

happy group or a sad group, doing a mood induction separately for each group. Another researcher might also be interested in the effects of mood on memory for an ad and decide to measure the naturally occurring mood of participants before they watch the ads. Because the IV is measured rather than manipulated in the second example, the method is a quasi-experiment, rather than a randomized (or true) experiment. There are sophisticated analytic tools for dealing with continuous IVs (e.g., see Hayes 2013), but measured IVs always have a higher likelihood of underlying systematic differences or confounding variables than do carefully controlled, manipulated IVs to which participants are randomly assigned. As such, it is more difficult to infer causal relationships with continuous predictor variables than with true, intentionally manipulated IVs.

Dependent variables may also be either discrete, such as choosing one brand over another, or continuous, such as a liking rating on a 7-point scale. To illustrate some of these ideas, consider an experiment in which some participants were repeatedly subliminally exposed to the brand name "Lipton Ice," a moderately popular canned tea, while control participants were subliminally exposed to "Nipeic Tol," a nonsense anagram of the brand (Karremans, Stroebe, and Claus 2006). All participants were told that they were engaged in a visual detection task, and—as a manipulation check on the efficacy of the subliminal stimuli—no one in either group reported seeing anything unusual. The dichotomous DV was whether participants chose Lipton Ice or another moderately popular beverage after the "visual detection task." Participants were more likely to choose Lipton Ice when subliminally presented with "Lipton Ice," but only if they also reported being thirsty. In this first study the authors used thirst as a continuous predictor variable, but because the authors measured, rather than manipulated, thirst, they were careful to replicate their finding in a second study that treated thirst as a true IV. In the second study the researchers controlled and manipulated participants' thirst by serving them a salty treat (known to induce thirst) before beginning the subliminal message portion of the experiment. Results of the second study were consistent with the first. Thus, the researchers could rule out the possibility that a hidden, third variable was causing both the thirst and subsequent tea choice.

Many advertising studies use DVs that are continuous, as in studies examining attitude change. Attitudes can be measured in different ways, but one common method is the semantic differential format (Osgood, Suci, and Tannenbaum 1957; Crites, Fabrigar, and Petty 1994) in which respondents are asked to assess a concept via a scale of polar opposite adjectives (e.g., *Good/Bad*). DVs can also be continuous ratio level, as when participants are asked to estimate the value of an item in dollars and cents, or when they are asked to allocate points to different response options. However, whether variables are discrete or continuous, manipulated or measured, it is critical that they accurately represent the concepts they are intended to represent.

## CONCEPTUALIZING AND OPERATIONALIZING VARIABLES

Variables manipulated and measured in experiments are often intended to represent higher-order constructs. Persuasion knowledge (Ham, Nelson, and Das 2015), advertising literacy (Nelson 2016), media multitasking (Chinchanchokchai, Duff, and Sar 2015), and humor (Yoon 2016) are examples of variables advertising researchers study. These variables are abstract concepts that can be measured and/or manipulated in a variety of ways. Moving from an abstract idea, or concept, to a manipulated or measured variable is the process of operationalizing that variable. By specifying concrete, directly observable operations researchers move from the subjective to the objective. Operationalizing variables is an essential part of the scientific process because “the operational definition removes the concept from the feelings and intuitions of a particular individual and allows it to be tested by anyone who can carry out the measurable operations” (Stanovich 2010, p. 39). Operationalization advances science by enabling researchers to test others’ work.

Operationalization of a variable begins with that variable’s conceptual definition, a verbal explanation of the meaning of a concept, expressing the central or core idea of the concept. A conceptual definition defines what the concept is, and what it is not. Advertising researchers commonly measure attitudes toward ads and brands. An attitude is not directly observable; it cannot be measured directly, like height or weight or temperature. Therefore, a conceptual definition of attitude is necessary.

One popular, contemporary conceptual definition of attitude is “a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor” (Eagly and Chaiken 1993, p. 1). This conceptual definition specifies that attitudes are mental constructs, existing as predispositions and manifesting as evaluations. This definition allows for a variety of operationalizations (see Cook and Seltiz 1964), ranging from overt self-report (e.g., giving a rating on Thurstone scales, semantic differentials, Likert scales; see Himmelfarb 1993 for precise explanations of these and other attitude measures) to performance on objective tasks (e.g., implicit association test; Greenwald, Nosek, and Banaji 2003), along with responses to “partially structured” stimuli (e.g., having people judge an ambiguous scenario, where their attitudes may influence their judgments, as in Vargas, von Hippel, and Petty 2004), and physiological responses, such as measuring activity in the facial muscles used for smiling and frowning (Cacioppo et al. 1986).

In some cases, one operationalization for a construct gains widespread acceptance and becomes used by a large number of researchers. One such example is attitude toward an ad ( $Att_{ad}$ ).  $Att_{ad}$  is defined as “a predisposition to respond in a favorable or unfavorable manner to a particular advertising stimulus during a particular exposure occasion” (MacKenzie and Lutz 1989, p. 49). Mitchell and Olson (1981) originally

identified four semantic differential pairings that loaded highly on an evaluative dimension of ratings for ads and used this to represent  $Att_{ad}$ . The four opposite pairs were *Good/Bad*; *Like/Dislike*; *Irritating/Not irritating*; and *Interesting/Uninteresting*. MacKenzie, Lutz, and Belch (1986) found that just two items (*Good/Bad* and *Like/Dislike*) did an adequate job of tapping this concept. Since that time, many advertising researchers have continued to use some subset of the items identified by Mitchell and Olson (1981) or have replaced some with very similar terms such as *Positive/Negative* (e.g., Halkias and Kokkinaki 2014).

Typically, well-established measures will have the advantage of being accepted by the field and often have already been shown to have acceptable reliability and validity. However, it is always important to think about whether the standard operationalization adequately matches the concept before using it.

## CONSTRUCT VALIDITY

When operationalizing IVs and DVs researchers must be concerned with the ability to generalize from the specific operationalizations in studies to the broader abstract concepts they are intended to represent. This concern is known as construct validity. Some of the most frequently mentioned problems and leading causes for manuscript rejection in journals involve problems with construct validity. Shadish, Cook, and Campbell (2002) have identified 14 distinct threats to construct validity. We review some of those threats that are most likely to affect advertising researchers here, but we urge interested readers to review the full list of threats to construct validity in Shadish, Cook, and Campbell’s (2002) excellent book.

### Inadequate Explication of Constructs

In any experiment, the primary variables must be explicitly defined and the measures/manipulations must then be derived from those definitions. Constructs may be described too generally, too narrowly, inaccurately, and/or imprecisely. Problems can come from poor definition of the concept, as well as poor measurement of that concept. For example, if a researcher wants to look at how positive feelings affect ad perception, she would need to consider whether she is more interested in something like positive mood or if she is interested in an emotional reaction. Cause, duration, and effects would differ depending on this, so specifying is important in terms of predictions, measures, and manipulations, as well as the ability of future researchers to accurately build on or replicate published work. Similarly, measurement or manipulation of the concept should be driven by the conceptual definition.

Calling a person happy simply because she smiles might be overgeneralizing why people smile. People smile in job interviews because they want to make a good impression, not necessarily because they are happy. A narrow explication might

involve arguing that happy mood induced by autobiographical recall represents all forms of happiness. Happiness induced by a pleasant memory is surely different from happiness induced by winning a megalottery. An inaccurate explication would be measuring happiness and calling it attitude. An imprecise explication would be labeling a self-report mood measure as simply a mood measure. Situations like these, where the operational definition fails to precisely match the conceptual or theoretical definition, are a major problem in manuscripts. In addition to ensuring that conceptual and operational definitions are connected and precise, pretesting the validity of constructs can be very helpful. For example, Zhao, Muehling, and Kareklas (2014) employed a neutral condition in a pretest to their second study to ensure that happy and sad mood manipulations truly reflected happy and sad moods and were not just statistically different from each other.

It might seem that questionable operationalizations and inadequate concept explications are easy to avoid, but the authors of this article have all frequently seen problems like this when serving as reviewers. For example, people measure attitude toward a brand when theoretically they expect attitude toward an ad should be affected (or vice versa). Or a theory about mood is used but immediate emotional response is manipulated. Similarly, researchers may use one type of memory measure, such as recognition, when really the theory would suggest a different measure, such as recall, is what would be affected.

Addressing these threats to construct validity requires careful thought and consideration among members of the research team and beyond. In developing a research study, a review of previous research to see how others have operationalized constructs can be most informative. However, researchers must still think carefully about previous measures and determine if they adequately reflect the conceptual definition of the variable. In fact, mixed results between studies may occur because what is being called the same concept is being operationalized in ways that reflect different concepts. In writing a manuscript, it is essential that the meaning and measurement of all key constructs be fully and accurately described. This is frequently an area of concern raised by manuscript reviewers.

### Construct Confounding

At the heart of strong experimentation is controlling as many variables as possible while only changing the variable being studied. Manipulating mood by showing participants assigned to the happy condition a video of a peaceful sunset, and people in the sad condition a video of a cheetah chasing and killing a baby giraffe confounds valence and arousal. The happy condition participants might indeed be happy, but due to the video they might also be calm. Participants in the sad condition might be sad and agitated. Thus, any differences in the groups could not be attributed to valence of the mood alone (negative or positive) but also possibly due to differences in

arousal. Confound and manipulation checks can be important ways to protect against confound threats to construct validity (see Perdue and Summers 1986 on the use of confound and manipulation checks), but researchers should always carefully think about controlling as much as possible while trying to isolate the variable of interest.

### Mono-Operation Bias

In many research papers constructs are operationalized identically across multiple studies. Any single operation is likely to underrepresent the full construct and contain specific irrelevancies that could affect results. Exact replications of studies are important because they increase confidence in findings, but in terms of construct validity it is better to use multiple operationalizations across studies. For example, Elder and Krishna (2010), in a three-study article on the effects of single versus multiple senses appeals on taste preferences, manipulated the appeals by using three-word taglines for gum in their first study; longer descriptive paragraphs for potato chips in their second study, and a different descriptive paragraph for popcorn in their third study. Using these different manipulations increases their construct validity, although they still may not be able to generalize beyond relatively brief statements and snack foods.

### Mono-Method Bias

As with mono-operation bias, using a single method in research limits generalizability. If multiple studies are reported in a manuscript, it is a wise idea to include different methods to operationalize key constructs to avoid any potential mono-method bias. Zhao, Muehling, and Kareklas (2014), for example, used two different methods to show that consumers' affective state moderated the effectiveness of nostalgic advertising. In one, affective state was primed by having participants recall positive (or negative) experiences from their own lives, while in the second they read a news story previously found to induce a happy (or sad) mood.

### Confounding Constructs With Levels of Constructs

Treatments administered to participants may be too weak to effect changes in DVs. For example, researchers comparing the effect of argument quality (e.g., strong versus weak arguments) on attitude change may discover no differences between the groups and conclude that argument quality has no effect on attitudes. If the argument quality manipulations are weak, or mild, the lack of difference may be due to the researchers not developing adequately strong and weak arguments. Advertising is often for brands that do not differ all that much, so message arguments in existing ads do not tend to be overly strong or weak. As a result, many advertising studies use manipulations that are rated very close to the midpoint of

scales of argument strength. This may reflect reality but can limit the ability to find differences or adequately test a theory under consideration. Pretesting manipulations with a group of respondents who will not be in the actual experiment can help ensure that manipulations adequately affect the concept of interest.

It is important to recognize that manipulations that result in significant differences between groups are not enough to ensure that researchers are adequately representing the intended concept. For example, a researcher may develop manipulations intended to represent high versus low involvement. A pretest using an involvement scale ranging from low involvement (1) to high involvement (7) shows that Manipulation A has a mean of 2.4 and Manipulation B a mean of 3.8. Even if the differences between the groups show that B is significantly higher in involvement than A, it is inappropriate to label Manipulation B as “high involvement” because its mean is below the midpoint of the scale. To adequately represent constructs, the manipulations not only need to be different from each other (statistically significant), they also need to be absolutely different (on different sides of the scale midpoint). Once this is accomplished, it could be said that differences are likely attributable to high versus low involvement. However, it would still not be known if differences in outcomes are driven by being lower involvement or higher involvement (or both). A third group representing a more neutral or midpoint of involvement could help serve as a baseline for the other groups in this case.

### Treatment-Sensitive Factorial Structure

Changes in a DV due to an IV may not occur on all dimensions of a measure. The effect of a persuasive message on an attitude measure could be hidden if the measure is treated as a single, global attitude measure rather than separate measures of the affective and cognitive components of attitudes. For example, Gorn, Pham, and Sin (2001) showed that high-arousal ads showed directional effects of valence on attitudes more strongly for attitude measures that reference self (“I like the ad”) versus the stimulus (“The ad is likable”). If a researcher was looking at effects of arousal on ad attitude and unknowingly used an attitude measure with some items that referenced self and some that referenced stimuli, then they may unknowingly be diluting the ability to see the effects on attitude. Careful attention to the theoretical framework, conceptualizations, and operationalizations should help in understanding what specific changes might be expected.

### Reactivity to the Experimental Situation

It is natural for people (i.e., research participants) to try to make sense of the situation in which they find themselves. They may try to guess what the researcher is trying to do and “help” the researcher by responding as they believe the

researcher would like, rather than responding to the experimental situation “naturally,” as they might if they did not know they were in an experiment. This can be particularly problematic in advertising research where people are presented with a single ad and then asked to provide a response. In such situations people are likely to pay abnormally high attention to the ad and attempt to determine what has been manipulated and what the focus of the research is. There are numerous ways to try to minimize this reactivity, including by disguising the true nature of the study (such as by being told the purpose was to assess a video game when the actual purpose was to look at attention to ads in the game; e.g., Lee and Faber 2007); breaking up the experiment by assessing DVs at a later point in time or after the experiment is ostensibly over (such as contacting people who were shown an ad 48 hours later to assess unaided recall of the ad; e.g., Friedman and Friedman 1979); and by using two different experimenters who claim to be studying different things, one presenting the IV and the other collecting the DV (e.g., Feinberg 1986). Rosenthal and Rosnow (2007) provide a more complete discussion of techniques for minimizing reactivity in various situations.

### Experimenter Expectancies

Just as teachers’ expectations for students can become self-fulfilling prophecies (Rosenthal and Jacobson 1968), experimenters’ expectations for research participants can threaten construct validity. Preventing experimenters from knowing which treatments participants receive can help reduce this problem, as can minimizing experimenters’ interactions with participants by, for example, presenting instructions to participants on a computer instead of face-to-face (Rosenthal and Rosnow 2007). In addition, when studies include subjective coded measures (e.g., coding open-ended recall of a commercial or consumer thought listings) it is important to demonstrate high agreement among multiple coders who are each unaware of (i.e., “blind” to) the conditions and hypotheses.

Other threats to construct validity exist, such as novelty and disruption effects (new treatments might generate excitement or disrupt a status quo), compensatory equalization (experimenters hoping to help research participants may offer some treatments to participants not assigned to that treatment), compensatory rivalry (participants in different groups compete to show they can do well), resentful demoralization (participants receiving an undesirable treatment may “give up” on participation), and treatment diffusion (participants in one condition receive the treatment intended for another condition). However, these threats are more likely to emerge in field experiments where participants receive some service or assistance program and can interact with one another to learn about different treatments (Shadish, Cook, and Campbell 2002). They seem less threatening to lab-based experimental research.

## INTERNAL VALIDITY

Internal validity refers to the extent to which researchers can be confident of a cause-and-effect relationship. An experiment with high internal validity allows greater certainty of cause and effect. Threats to internal validity are alternative explanations for observed findings, such as changes in the DV that may be due to events outside of the experiment, or research participants maturing during the study, becoming sensitized to the topic under study, or dropping out of the study unexpectedly. For example, if a researcher were to give study participants a small gift, such as a chocolate, to manipulate happiness, he would not know if effects were due to happiness from the gift or due to effects from the sugar in the chocolate. A full explanation of common threats to internal validity is beyond the scope of the present article, but we point interested readers to excellent work on this topic by Campbell and Stanley (1963) and Shadish, Cook, and Campbell (2002).

## EXTERNAL AND ECOLOGICAL VALIDITY

External validity usually refers to researchers' ability to generalize from the specific components of one particular experiment to other people, settings, treatments, and outcomes (Campbell and Stanley 1963; Cronbach et al. 1985; Shadish, Cook, and Campbell 2002). Stated more elegantly, "most experiments are highly local but have general aspirations" (Shadish, Cook, and Campbell 2002, p. 18). External validity is an issue in all types of experimental research but can be of particular concern in lab studies in applied fields such as advertising.

One major concern regarding external validity involves the participants used in an experiment. For at least the past 70 years (McNemar 1946), people have noted that a large number of academic lab studies use college students as participants. This has led to a debate of the generalizability of such research and questions about whether "college sophomores are really people" (Gordon, Slade, and Schmitt 1986; Greenberg 1987; Sears 1986). More recently, a similar concern has emerged over the limited demographic variability of most research participants. A large majority of academic research in advertising, marketing, consumer behavior, psychology, and behavioral economics is conducted with people who are identified by the acronym WEIRD (Western, Educated, from Industrialized countries, relatively Rich, and from Democratic countries; Henrich, Heine, and Norenzayan 2010). This narrow, nonrepresentative slice of humanity almost certainly prevents researchers from generalizing experimental findings to most of the other people in the world, much less other settings, treatments, and outcomes.

In some situations this may not create a problem with external validity, but in others it may. Nearly two-thirds of the world's population lives in Asia, which tends to be more collectivistic compared to individualistic Western culture. Collectivistic cultures prioritize group harmony, whereas

individualistic cultures prioritize autonomy; and effective advertising in Korea and the United States reflects these different priorities. Accordingly, Han and Shavitt (1994) found that U.S. participants preferred individualistic ads (e.g., "Treat yourself to a breath freshening experience," p. 336), while Korean participants preferred collectivistic ads (e.g., "Share the Freedent breath freshening experience," p. 336). But imagine another pair of researchers studying the efficacy of personal appeals versus group appeals in advertisements, and these researchers collect data only in the United States. They would have found that personal appeals caused more favorable attitudes than group appeals, and they might be tempted to generalize their findings globally. Thanks to Han and Shavitt (1994) we know the impact of individual versus group appeals is not the same in all cultures.

Threats to external validity can also emerge from the specific stimuli researchers use and the setting in which it is studied (e.g., laboratory, store, home, car) (Shadish, Cook, and Campbell 2002). Will a finding hold for different ads, different products, or different media (e.g., print ad, video ad, pop-up ad online, banner ad)? Will people respond similarly to a commercial presented in a laboratory setting as they would to that ad appearing on their television at home? Unfortunately, it is impossible for any one study to include all possible types of stimuli and locations. To some extent, researchers rely on replications using different types of participants, stimuli, media, and settings to demonstrate if a theory can be generalized beyond the limits of its initial testing. This is why theory testing is a cumulative process that takes place over time and why the results of any one study do not "prove" a theory but merely demonstrate support for it. However, this does not mean that it is a worthwhile endeavor to take a particular theory and continually replicate it using different participants, media, or stimuli. Rather, researchers must use logic and reasoning to determine when and why a theory may not hold. The noted example of collectivist versus individualistic culture is a good example of where a theory allowed researchers to predict previously unexpected differences between groups.

Ecological validity refers to whether the elements of experiments are consistent with the types of things people encounter in everyday life. Advertisements shown on prime-time network and cable TV usually have high production values and cost tens, if not hundreds, of thousands of dollars to make. Ads typically used in experiments have only slight resemblance to the ones seen outside of the laboratory. Similarly, in real life people often see the same ad numerous times and frequently have existing beliefs and attitudes toward the brand being advertised prior to seeing the ad. Researchers often have to make trade-offs between maximizing ecological validity and maximizing control of the variables of interest and eliminating potential confounds.

In some situations external and ecological validity are relatively unimportant in the context of the experimental research. Most academic research uses experimental studies to test



causal hypotheses; academic researchers typically do not conduct experiments to determine the likelihood that some event will occur in a particular population, with particular treatments, in particular outcomes, or in particular settings (Berkowitz and Donnerstein 1982). This is a major difference between academic research and industry research. The scientific study of advertising does not always require that researchers try to meticulously re-create the everyday experience of encountering advertising in the real world. Experimental research is sometimes conducted by prying variables apart, and even by sometimes intentionally creating unrealistic stimuli and situations, precisely so researchers can control specific variables, test hypotheses, and understand cause-and-effect relationships (Banaji and Crowder 1989). For example, if a researcher wants to understand the impact of humor on attention to an ad, she may want to carefully control factors in a lab and use a single exposure to a stimulus. However, if the researcher's theory of interest is the wear-out effect of humor on attention, then she would need to include multiple exposures of the ad. Theory should dictate the way researchers conduct their experiments and the type of threats they want to eliminate. This should guide researchers in making experimental design decisions, such as the appropriateness of using single versus multiple ad exposures; real versus unknown or made-up brands; and when it is important to have professionally created ads or simply test a concept using artificially created ads.

## EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS

There are many ways to conduct experimental research, and they adhere to a few different basic experimental designs (for a complete discussion of these designs, refer to Campbell and Stanley 1963; Shadish, Cook, and Campbell 2002). These designs can be divided into two major categories: quasi-experiments and randomized experiments. Quasi-experimental designs lack at least one of the two critical features of "true" randomized experiments: (1) control or comparison groups and (2) random assignment to groups. Quasi-experiments are generally susceptible to threats to internal validity and are therefore poor tools for researchers hoping to establish cause-and-effect relationships; however, sometimes they are necessary when researchers cannot practically or ethically manipulate some variable, such as respondents' cultural background (e.g., Han and Shavitt 1994) or a personality trait.

### Quasi-Experimental Designs

*One shot.* In this simplest quasi-experimental design, researchers administer a treatment (X) to a single group, followed by observation (O). For example, a researcher might show an ad to participants and then ask them how much they like the brand featured in the ad. The problem here is that the researcher cannot know if seeing the ad made any difference

on participants' liking of the brand. To overcome this limitation, researchers need some form of control group with which to compare the results.

*One group pretest-posttest.* One way to do this is to compare brand attitudes of each person before and after they see the ad. Here researchers would make an observation of a single group ( $O_1$ ), administer a treatment (X), and then make another observation of that same group ( $O_2$ ). In this case, the difference between the group at Time 2 compared to Time 1 ( $O_2 - O_1$ ) is the effect of X. However, this design is susceptible to many threats to internal validity. For example, between the pretest and posttest, other events (e.g., negative news about the brand) could have influenced the change (a history threat to internal validity). Because one cannot rule out threats to internal validity, causal relationships cannot be established with any certainty.

*Static group.* Another way to introduce a control group is to have two different groups that are compared. One group receives some treatment ( $O_{\text{treat}}$ ) while the other group does not ( $O_{\text{ctrl}}$ ). The same instrument is used to make observations of the two different groups, and the difference between the two groups ( $O_{\text{treat}} - O_{\text{ctrl}}$ ) is the effect of X. For example, researchers may wonder whether seeing a movie that includes a prominent brand placement (e.g., a Ducati motorcycle) influences attitudes toward that brand. To determine this the researchers could ask people whether they saw the movie and then ask their evaluation of a few brands, including Ducati. The researchers could then determine if people who saw the movie had a more positive attitude toward the motorcycle brand than those who did not to see it. However, the researchers still would not be able to rule out other factors, such as systematic difference in group membership, as the cause of any differences. For example, if it was an action movie, people who saw it may be higher in sensation seeking than the nonviewers; and high sensation seekers may be more favorable toward motorcycles in general.

### Random Assignment

Randomly assigning participants to different treatment conditions serves to overcome some of the concerns in the previous designs. With sufficiently large samples, randomly assigning people to different groups makes the groups, on average, nearly identical to one another. Happy and sad people, rich and poor alike, have equal chances of being in the experimental and control groups. The same goes for any characteristic on which people vary: intelligence, gender, age, sense of humor, mood, extraversion, motivation, and so on. Thus, the experimental and control groups should be very nearly identical, on average, before they are exposed to the experimental treatment, and any difference that is observed is most likely due to the treatment.

Random assignment is not perfect, so it does not guarantee that the experimental and control groups are identical. But it is

a remarkably simple and effective way to deal with the amazing variability among people. Random assignment is less effective in equalizing groups when there are fewer participants in a study because small samples will vary more. For example, consider a study with only four participants: two males (M) and two females (F). If participants are randomly assigned to each condition, the only possible outcomes are MM, MF, FM, or FF. There is a 50% likelihood that one condition will have only male participants and the other will have only females. If this occurs, it will be impossible to know if any difference observed is due to the treatment or the sex of the respondents.

Now consider a study with 100 participants, 50 males and 50 females. The likelihood of randomly assigning participants to conditions such that all men end up in one condition and all women end up in the other would be extremely small. It is certainly possible that random assignment could result in groups with, say, 23 women in one group and 27 women in another, but this is a relatively small difference, unlikely to create a plausible alternative explanation for an observed difference following a treatment. By adding randomization and control groups to the last two quasi-experiments, researchers can turn them into experimental designs.

### Randomized Experiments

There are several randomized experimental designs, including a before-and-after with control design that builds on the pre-post quasi-experimental design by adding a control group and randomization, and the Solomon four-group design that combines the before-and-after with control design with an after-only with control and randomly assigns participants to all four conditions. However, by far the most common design used in advertising research is the

after-only with control. The after-only with control design adds the randomization of participant assignment to groups to the quasi-experimental static group design. Randomly assigning participants to either a group that receives the treatment or one that does not (control group) allows researchers to infer cause-and-effect relationships quite well.

### Common Issues in Advertising Experiments

It has been our experience that the most common study design in the advertising literature involves (1) an experimental design of after-only with control (here “control” refers to a group that receives a different treatment, not necessarily a group that receives no treatment) (2) using a student sample (3) shown ads for products that students are typically familiar with, (4) using hypothetical or unknown brands, (5) with only a single exposure to the ad, (6) in situations of forced exposure. While each of these characteristics may well be appropriate in some situations, there are also many times when they are not.

As previously explained, the after-only with control is a very useful experimental design that relies on randomization to help rule out many threats to validity. However, the focus of the research—the theoretical framework and concepts—should dictate the design of the study. Other randomized designs may be more appropriate for some research questions. For example, to see the impact of an opt-in e-mail marketing campaign, DuFrene and colleagues (2005) had the same people indicate their brand attitudes, purchase intent, and feelings of trust toward a company both before and after receiving differing numbers of e-mails from the company. In situations where researchers are testing effects of an ongoing ad campaign or looking for differences in immediate and delayed responses to ads (e.g., Lariscy and Tinkham 1999) a before-and-after with control design may be more appropriate to test the theoretical ideas.

Scholars have debated the use of student samples for years. However, students remain frequent experimental participants because their use has numerous benefits. These include low costs, the ease of obtaining participants, the potential ease of getting them to the experimental setting, and the fact that the homogeneity of students makes it easier to find products and ad appeals that are appropriate for all the participants. However, college students are not always appropriate for the theory being tested. Studies looking at the development of an understanding of what a commercial is (e.g., Ward, Wackman, and Wartella 1977) or the development of advertising literacy and persuasion knowledge (e.g., Nelson 2016) or examinations of differences across the life span (e.g., Stephens 1982) clearly require non-college students. Even when college students may be age appropriate, it is critical to consider if they are theoretically appropriate for the question researchers want to address.

Meaningful research in advertising requires the stimuli, such as the products used in experiments, to match the theory being tested. Many research studies in advertising are interested in involvement and central versus peripheral routes to persuasion. Several of these studies use expensive cars to represent a high-involvement product. However, is a BMW or Mercedes a high-involvement product to most college students? It is hard to imagine that many of them would be able to spend \$50,000 to \$100,000 on a new car. A pretest might be conducted that shows students believe BMWs are expensive and require research and comparison before purchase. However, if they are not currently in the market for a car, and the outcome is not actually purchasing a car, then they are unlikely to have the intrinsic motivation that an actual car purchase might create. Are they, then, truly motivated and capable of processing the message for a new luxury car, and can they really assess the likelihood they would purchase such a car? It is important to evaluate the product and brands used in a study in the context of how respondents will actually process messages for these products and how this might affect the primary variables in the study.

In addition, it is critical to clearly define concepts and derive operationalizations from those conceptual definitions. For example, a researcher might define involvement in a study as a “purchase decision that has high personal relevance and importance” (e.g., Zaichkowsky 1986). He then would need to ensure that he is using materials that are high versus low in involvement to the population from which he is sampling. Thus, if participants are students, the researcher would need to pretest products/brands that they currently purchase and consider to be more (less) relevant and important to them.

Many advertising experiments use made-up or relatively unknown brands. This has the advantage of ensuring that previous knowledge, attitudes, or experiences with the brand do not influence the study results. However, this also creates a problem of relevance of the findings to the typical advertising situation. Most advertising is for brands people already know. Are people equally likely to attend to, and recall, ads for both known and unknown brands? If not, the choice of a real or hypothetical brand may influence the study results. Most of what we know from ad research may tell us more about how advertising affects the introduction of new brands than the clear majority of advertising that is for existing brands. Researchers need to think about this, and similar issues, in designing experiments. These trade-offs often end up as single lines in limitation sections of manuscripts, but it is critically important to weigh the costs and benefits before designing the study, procuring the materials, and conducting the experiment. Simply because a published study was performed a certain way does not mean that same procedure is right for another study.

Similarly, the majority of advertising studies rely on a single exposure to an ad. This can be useful for testing some theories, but it is also clear that multiple exposures may be different than the initial exposure to an ad (Cacioppo and Petty 1980). In addition, some theories (such as nonevaluative associative learning) assume that a single exposure may not be sufficient to produce an effect, but multiple exposures will be (e.g., Lutchyn and Faber 2016). In situations like these, it is important to have a design that allows for ad repetition. If such a design is desired, then additional decisions, such as whether to use massed (multiple repetitions all occurring in a single exposure setting) or spaced distribution of messages (repeated exposures spread out over time), need to be determined (for more information about ad repetition, see Pechmann and Stewart 1988).

Finally, most advertising experiments are designed to focus the participants’ attention on the advertising stimulus. People are asked to look at the ad or are shown a screen containing just the advertising message. This ensures attention to the message and increases the likelihood that a possible effect will occur. However, most people do not view ads this way in their daily lives. Instead, they see ads in a cluttered environment while engaged in other activities and offer those ads only partial attention at best. Ads designed to break through clutter and

gain attention may demonstrate greater effectiveness in situations where attention is not demanded rather than when it is required as part of the experiment. Advertising studies concerned with attention most often require nonforced situations to be meaningful. The directions given to participants can also impact the findings of a study. For example, Duff (2009) found that directions to browse a website versus looking for specific content information leads to very different effects on peripheral ads.

Of course, there is no one correct design for advertising experiments. Each individual decision made in designing a study requires some in-depth thinking regarding how it fits with the theory being tested. It is the theory that should drive the experimental design.

### More Complex Experimental Designs

To this point we have reviewed only those experimental designs intended to test the effect of one IV on one or more DVs. These are known as one-way, between-participants designs because they test only one IV and because participants are assigned to one, and only one, treatment condition, and the comparisons are between the groups. There are many other types of designs, but due to space limitations we are unable to go into their details here. There are within-participants, or repeated measures, experimental designs, where the same participants may be exposed to more than one treatment (see, e.g., Peterson et al. 2016). For example, participants may see an ad with a celebrity endorser and then see a different ad with a typical consumer endorser, or vice versa. Between-participants designs, where participants see only one treatment condition, are used more often in advertising research. There are several reasons for this, but one important reason is that between-participants designs make it more difficult for participants to guess what the study is about. Participants who see a celebrity endorser and then a typical consumer endorser may realize that the study is about endorsers or message source, and respond how they think the researcher wants them to respond (known as a demand effect; see Shimp, Hyatt, and Snyder 1991) rather than how they would naturally respond.

In addition, we have considered only one-way designs that have two groups: an experimental group and a control group, or two different conditions. A one-way design could have more than two levels; researchers may want to study the effects of one IV with three or more different levels. For example, a researcher might want to investigate the effects of repetition on attitudes toward an ad, so she could randomly assign participants to different levels of repetition, where different people see an ad one, five, or 10 times before rating it.

We have focused on the most basic type of experiment with one IV and one DV. Often, however, researchers are interested in the effects of two or more IVs on one (or more) DV. A two-way factorial design involves two IVs combined in all possible

ways. For example, recall the subliminal message study (Karremans, Stroebe, and Claus 2006) reviewed briefly earlier in this article. One IV in that study was the type of stimuli presented subliminally: the name of a real brand of tea or a nonsense word. The other IV in that study was participants' thirst level: half of the participants were given a salty treat to make them thirsty and half were not. In this example, there are two IVs (type of stimulus and degree of thirst), each with two levels (type of stimulus: brand name versus nonsense word; degree of thirst: high versus low). This is a two-way factorial design because there are two IVs, combined in all possible ways to create four different treatment conditions, with participants randomly assigned to one, and only one, of the four conditions: brand + high thirst; brand + low thirst; nonsense word + high thirst; nonsense word + low thirst. This design could also be called a  $2 \times 2$  factorial, where the first 2 refers to an IV with two levels, and the second 2 refers to a second IV, also with two levels. A  $2 \times 3$  factorial design would have two IVs, one with two levels and the other with three levels, for a total of six possible treatment conditions. A  $2 \times 2 \times 2$  design would have three IVs, each with two levels, for a total of eight conditions.

In factorial experiments there may be multiple outcomes of interest. Consider a  $2 \times 2$  factorial experiment testing the effects of argument quality (strong versus weak) and personal involvement (high versus low) on attitudes toward a persuasive communication (Petty, Cacioppo, and Schumann 1983). There may be an effect of argument quality on attitudes, such that, on average, participants who read a message with strong arguments will have more favorable attitudes than those who read a message with weak arguments. This would be a main effect of argument quality. There could also be a main effect of personal involvement (e.g., participants who were highly involved have more favorable attitudes than those who were not involved). For this example, suppose there is no main effect of involvement. However, involvement may still have an effect on attitudes. There may be an interaction effect, where the effect of one IV on the DV depends on the level of the other IV. In this example, the effect of argument quality on attitudes might depend on the level of personal involvement: Participants who were highly involved showed a larger effect of argument quality (involved participants who read strong arguments were persuaded in favor of the message, while involved participants who read weak arguments were persuaded against the message), while participants who were less involved showed no effect of argument quality (they had more neutral attitudes regardless of whether they read strong or weak arguments).

For a two-way factorial experiment, any combination of two main effects and one interaction may (or may not) emerge. For a three-way factorial experiment (e.g., three IVs that we will label A, B, and C) researchers have to consider three possible main effects (one for each IV), three possible two-way interactions ( $A \times B$ ,  $A \times C$ , and  $B \times C$ ), and one possible

three-way interaction ( $A \times B \times C$ ). For a four-way factorial, there may be four main effects, six possible two-way interactions, three possible three-way interactions, and one possible four-way interaction. Factorial design experiments tend to become quite difficult to interpret when there are more than three IVs, and the interpretation of their results can become very complex and rather messy. Readers interested in going into more depth on this issue should consult experimental design and statistics resources such as those introduced by Iacobucci (2001), Iacobucci and Churchill (2009), or Rosenthal and Rosnow (2007).

## ETHICAL ISSUES FOR PARTICIPANTS AND RESEARCHERS

Researchers have a responsibility to treat research participants ethically. Universities almost always have ethics review boards who examine proposed research to ensure minimal harm to participants before the research can begin. This involves three main principles: respect for people, beneficence, and justice (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979). Respect for people means treating people as autonomous agents and protecting those who may have impaired autonomy. Therefore, participants must provide informed consent to participate in university experiments. Beneficence means not harming people and maximizing benefit for them. Participating in experiments should involve minimal risk and have some benefit for participants, such as educational benefit or remuneration for participation. Justice means we must consider who benefits from research and who may be harmed by it.

Many experiments involve some form of deception (e.g., in a between-participants experiment not informing people of the existence of other experimental conditions). A review of published marketing research covering three decades (1975–76, 1989–90, and 1996–97) revealed that 43.4% to 58.5% of studies involved the use of deception (Kimmel 2001). Researchers have an obligation to inform participants about the true nature of the study once they are no longer involved as participants. Even though passive deception (i.e., not informing participants about other conditions) may seem a mild transgression, there are those who believe that deceptive research is never justified because it can harm participants, the profession, and even society (Baumrind 1985).


Researchers also have a responsibility to conduct and report work ethically, being honest about findings or lack thereof. Obviously, researchers should never fabricate data or results, but experiments can be conducted, analyzed, and reported in ways that maximize the likelihood of obtaining significant (i.e., publishable) but questionable results. This too can be an ethical issue.

For example, in one published experiment, 20 participants were randomly assigned to listen to either "When I'm 64" by

The Beatles or “Kalimba,” a song that is part of the Windows 7 computer operating system, and then indicate their age. Their findings showed that listening to “When I’m 64” caused people to be older, compared to those assigned to listen to “Kalimba” (Simmons, Nelson, and Simonsohn 2011). However, the real purpose of their research was to highlight how “acceptable” experimental practices can be used to obtain false-positive results. They obtained their results by (1) choosing their sample size for the experiment by stopping data collection as soon as they obtained significant results, (2) choosing only a few selected DVs to report from among many they collected, (3) using covariates to analyze their data, and (4) failing to report all the experimental conditions they tested. In computer simulations Simmons, Nelson, and Simonsohn (2011) demonstrated that using all four of these techniques increased the likelihood of obtaining an incorrect significant finding (i.e., a false positive) from the conventional 5% ( $p < .05$ ) to 60.7%. This has the potential to harm the field (by creating a false knowledge base), the individual researcher’s reputation (by publishing studies that are false and cannot be replicated), the participants (whose time is wasted by participating in illegitimate research), and society (who may believe and use information from false-positive research). Researchers have a responsibility to all of these groups to conduct and report research honestly by deciding on sample sizes before conducting research, using appropriately large samples, reporting all variables and conditions in studies, reporting results with and without covariates, not reporting only those studies that found significant results while excluding ones that did not, and reporting results with and without observations that are deemed outliers or otherwise unsuitable (see Simmons, Nelson, and Simonsohn 2011).

Experiments can be powerful tools for increasing understanding of why things occur and helping predict when something might happen again. In advertising research, it is important to understand the trade-offs involved in designing an experiment. While advertising is an applied field, there is still a need to use carefully developed variables and well-controlled designs guided by theory and careful thought. In this way researchers can better hope to know when something might lead to a specific outcome. However, in some cases, settings or stimuli that are too controlled could compromise validity and potentially threaten the predictive power that is gained. For example, the forced attention to an ad that occurs in many lab settings may not lead to the same effects as partial attention to an ad occurring in more natural situations. The goal of the research should help us determine which trade-offs are more crucial in any particular study. The key to using experiments to improve the quality of advertising theory and better understand effects is to (1) develop studies that are guided in their design by theory and logic, (2) strive to control for, or rule out, alternative explanations, and (3) maintain high ethical standards.

## ORCID

Brittany R. L. Duff  <http://orcid.org/0000-0002-3206-0353>

## REFERENCES

- Banaji, Mahzarin R., and Robert G. Crowder (1989), “The Bankruptcy of Everyday Memory,” *American Psychologist*, 44 (9), 1185–93.
- Baumrind, Diana (1985), “Research Using Intentional Deception: Ethical Issues Revisited,” *American Psychologist*, 40 (2), 165–74.
- Berkowitz, Leonard, and Edward Donnerstein (1982), “External Validity Is More Than Skin Deep: Some Answers to Criticisms of Laboratory Experiments,” *American Psychologist*, 37 (3), 245–57.
- Bhat, Subodh, Thomas W. Leigh, and Daniel L. Wardlow (1998), “The Effect of Consumer Prejudices on Ad Processing: Heterosexual Consumers’ Responses to Homosexual Imagery in Ads,” *Journal of Advertising*, 27 (4), 9–28.
- Caballero, Marjorie J., and Paul J. Solomon (1984), “Effects of Model Attractiveness on Sales Response,” *Journal of Advertising*, 13 (1), 17–23, 33.
- Cacioppo, John T., and Richard E. Petty (1980), “Persuasiveness of Communications Is Affected by Exposure Frequency and Message Quality: A Theoretical and Empirical Analysis of Persisting Attitude Change,” *Current Issues and Research in Advertising*, 3 (1), 97–122.
- , Mary E. Losch, and Hai Sook Kim (1986), “Electromyographic Activity over Facial Muscle Regions Can Differentiate the Valence and Intensity of Affective Reactions,” *Journal of Personality and Social Psychology*, 50 (2), 260–68.
- Campbell, Donald T., and Julian C. Stanley (1963), “Experimental and Quasi-Experimental Designs for Research on Teaching,” in *Handbook of Research on Teaching: A Project of the American Educational Research Association*, N.L. Gage, ed., Chicago: Rand McNally, 171–246.
- Chang, Chun-Tuan, and Ching-Ting Yen (2013), “Missing Ingredients in Metaphor Advertising: The Right Formula of Metaphor Type, Product Type, and Need for Cognition,” *Journal of Advertising*, 42 (1), 80–94.
- Chinchanachokchai, Sydney, Brittany R. L. Duff, and Sela Sar (2015), “The Effect of Multitasking on Time Perception, Enjoyment, and Ad Evaluation,” *Computers in Human Behavior*, 45, 185–91.
- Cline, Thomas W., and James J. Kellaris (2007), “The Influence of Humor Strength and Humor-Message Relatedness on Ad Memorability: A Dual Process Model,” *Journal of Advertising*, 36 (1), 55–67.
- Cook, Stuart W., and Claire Selltiz (1964), “A Multiple-Indicator Approach to Attitude Measurement,” *Psychological Bulletin*, 62 (1), 36–55.
- Crites, Stephen L., Leandre R. Fabrigar, and Richard E. Petty (1994), “Measuring the Affective and Cognitive Properties of Attitudes: Conceptual and Methodological Issues,” *Personality and Social Psychology Bulletin*, 20 (6), 619–34.
- Cronbach, Lee J., Sueann Robinson Ambron, Sanford M. Dornbusch, Robert D. Hess, Robert C. Hornik, D.C. Phillips, Decker F. Walker, and Stephen S. Weiner (1985), *Toward Reform of Program Evaluation*, San Francisco, CA: Jossey-Bass.
- Duff, Brittany R.L. (2009), “The Eye of the Beholder: Affective and Attentional Outcomes of Selective Attention to Advertising,” doctoral dissertation, University of Minnesota.
- DuFrene, Debbie D., Brian T. Engelland, Carol M. Lehman, and Rodney A. Pearson (2005), “Changes in Consumer Attitudes Resulting from Participation in a Permission E-Mail Campaign,” *Journal of Current Issues and Research in Advertising*, 27 (1), 65–77.
- Eagly, Alice H., and Shelly Chaiken (1993), *The Psychology of Attitudes*, San Diego, CA: Harcourt Brace Jovanovich College Publishers.
- Elder, Ryan S., and Aradhna Krishna (2010), “The Effects of Advertising Copy on Sensory Thoughts and Perceived Taste,” *Journal of Consumer Research*, 36 (5), 748–56.
- Elmore, Joann G., and Alvan R. Feinstein (1994), “Joseph Goldberger: An Unsung Hero of American Clinical Epidemiology,” *Annals of Internal Medicine*, 121 (5), 372–75.

- Feinberg, Richard A. (1986), "Credit Cards as Spending Facilitating Stimuli: A Conditioning Interpretation," *Journal of Consumer Research*, 13 (3), 348–56.
- Friedman, Hershey H., and Linda Friedman (1979), "Endorser Effectiveness by Product Type," *Journal of Advertising Research*, 19 (5), 63–71.
- Gordon, Michael E., L. Allen Slade, and Neal Schmitt (1986), "The 'Science of the Sophomore' Revisited: From Conjecture to Empiricism," *Academy of Management Review*, 11 (1), 191–207.
- Gorn, Gerald, Michel T. Pham, and Leo Yatming Sin (2001), "When Arousal Influences Ad Evaluation and Valence Does Not (and Vice Versa)," *Journal of Consumer Psychology*, 11 (1), 43–55.
- Greco, Alan J., and Linda E. Swayne (1992), "Sales Response of Elderly Consumers to Point-of-Purchase Advertising," *Journal of Advertising Research*, 32 (5), 43–53.
- Greenberg, Jerald (1987), "The College Sophomore as Guinea Pig: Setting the Record Straight," *Academy of Management Review*, 12 (1), 157–59.
- Greenwald, Anthony G. (2012), "There Is Nothing So Theoretical as a Good Method," *Perspectives on Psychological Science*, 7 (2), 99–108.
- , Brian A. Nosek, and Mahzarin R. Banaji (2003), "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm," *Journal of Personality and Social Psychology*, 85 (2), 197–216.
- Halkias, Georgios, and Flora Kokkinaki (2014), "The Degree of Ad-Brand Incongruity and the Distinction between Schema-Driven and Stimulus-Driven Attitudes," *Journal of Advertising*, 43 (4), 397–409.
- Ham, Chang-Dae, Michelle R. Nelson, and Susmita Das (2015), "How to Measure Persuasion Knowledge," *International Journal of Advertising*, 34 (1), 17–53.
- Han, Sang-Pil, and Sharon Shavitt (1994), "Persuasion and Culture: Advertising Appeals in Individualistic and Collectivistic Societies," *Journal of Experimental Social Psychology*, 30 (4), 326–50.
- Hayes, Andrew F. (2013), *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach*, New York: Guilford Press.
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010), "The Weirdest People in the World?," *Behavioral and Brain Sciences*, 33 (2–3), 61–83.
- Himmelfarb, Samuel (1993), "Attitude Measurement," in *The Psychology of Attitudes*, A.H. Eagly, and S. Chaiken, eds., Fort Worth, TX: Harcourt Brace Jovanovich, 23–87.
- Iacobucci, Dawn (2001), "Methodological and Statistical Concerns of the Experimental Behavioral Researcher: Introduction," *Journal of Consumer Psychology*, 10 (1), 1–2.
- , and Gilbert A. Churchill (2009), *Marketing Research: Methodological Foundations*, 10th ed., Mason, OH: South-Western College Publishing.
- Jacobellis v. Ohio, 378 U.S. 184 (1964).
- Kamins, Michael A. (1990), "An Investigation into the 'Match-up' Hypothesis in Celebrity Advertising: When Beauty May Be Only Skin Deep," *Journal of Advertising*, 19 (1), 4–13.
- Karremans, Johan C., Wolfgang Stroebe, and Jasper Claus (2006), "Beyond Vicary's Fantasies: The Impact of Subliminal Priming and Brand Choice," *Journal of Experimental Social Psychology*, 42 (6), 792–98.
- Kimmel, Allan J. (2001), "Deception in Marketing Research and Practice: An Introduction," *Psychology and Marketing*, 18 (7), 657–61.
- Lariscy, Ruth Ann Weaver, and Spencer F. Tinkham (1999), "The Sleeper Effect and Negative Political Advertising," *Journal of Advertising*, 28 (4), 13–30.
- LaTour, Kathryn A., and Michael S. LaTour (2009), "Positive Mood and Susceptibility to False Advertising," *Journal of Advertising*, 38 (3), 127–42.
- Lee, Mira, and Ronald J. Faber (2007), "The Effects of Brand Placement in Advergaming on Brand Memory: 'Let the Games Begin!'," *Journal of Advertising*, 36 (4), 75–90.
- Lutchyn, Yuliya A., and Ronald J. Faber (2016), "A New Look at Associative Learning in Advertising: Can Messages Influence Contextual Associations?," *Journal of Current Issues and Research in Advertising*, 37 (1), 28–44.
- MacKenzie, Scott B., and Richard J. Lutz (1989), "An Empirical Examination of the Structural Antecedents of Attitude toward the Ad in an Advertising Pretesting Context," *Journal of Marketing*, 53 (2), 48–65.
- , ———, and George E. Belch (1986), "The Role of Attitude toward the Ad as a Mediator of Advertising Effectiveness: A Test of Competing Explanations," *Journal of Marketing Research*, 23 (2), 130–43.
- McNemar, Quinn (1946), "Opinion-Attitude Methodology," *Psychological Bulletin*, 43 (4), 289–374.
- Mill, John S. (1843), *A System of Logic*, 2 vols., London: John W. Parker.
- Mitchell, Andrew A., and Jerry C. Olson (1981), "Are Product Attribute Beliefs the Only Mediator of Advertising Effects on Brand Attitude?," *Journal of Marketing Research*, 18 (3), 318–32.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979), *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Bethesda, MD: Department of Health, Education, and Welfare.
- Nelson, Michelle R. (2016), "Developing Persuasion Knowledge by Teaching Advertising Literacy in Primary School," *Journal of Advertising*, 45 (2), 169–82.
- Nisbett, Richard E., and Timothy D. Wilson (1977), "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, 84 (3), 231–59.
- Nordhielm, Christie L. (2002), "The Influence of Level of Processing on Advertising Repetition Effects," *Journal of Consumer Research*, 29 (3), 371–82.
- Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum (1957), *The Measurement of Meaning*, Urbana: University of Illinois Press.
- Pechmann, Cornelia, and David W. Stewart (1988), "Advertising Repetition: A Critical Review of Wear-In and Wear-Out," *Current Issues and Research in Advertising*, 11 (1–2), 285–330.
- Perdue, Barbara C., and John O. Summers (1986), "Checking the Success of Manipulations in Marketing Experiments," *Journal of Marketing Research*, 23 (4), 317–26.
- Peterson, Matthew, Kevin Wise, Yilin Ren, Zongyuan Wang, and Jiachen Yao (2016), "Memorable Metaphor: How Different Elements of Visual Rhetoric Affect Resource Allocation and Memory for Advertisements," *Journal of Current Issues and Research in Advertising*, 38, 65–74.
- Petty, Richard E., John T. Cacioppo, and David Schumann (1983), "Central and Peripheral Routes to Advertising Effectiveness: The Moderating Role of Involvement," *Journal of Consumer Research*, 10 (2), 135–46.
- Platt, John R. (1964), "Strong Inference," *Science*, 146 (3642), 347–53.
- Popper, Karl R. (1959), *The Logic of Scientific Discovery*, New York: Harper & Row.
- Pracejus, John W., G. Douglas Olsen, and Thomas C. O'Guinn (2006), "How Nothing Became Something: White Space, Rhetoric, History, and Meaning," *Journal of Consumer Research*, 33 (1), 82–90.
- Putrevu, Sanjay, and Kenneth R. Lord (1994), "Comparative and Noncomparative Advertising: Attitudinal Effects Under Cognitive and Affective Involvement Conditions," *Journal of Advertising*, 23 (2), 77–90.
- Rosenthal, Robert, and Lenore Jacobson (1968), "Pygmalion in the Classroom," *The Urban Review*, 3, 16–20.
- , and Ralph L. Rosnow (2007), *Essentials of Behavioral Research: Methods and Data Analysis*, 3rd ed., New York: McGraw-Hill.
- Sears, David O. (1986), "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature," *Journal of Personality and Social Psychology*, 51 (3), 515–30.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed., Boston, MA: Houghton Mifflin.
- Shapiro, Stewart, and Deborah J. MacInnis (2002), "Understanding Program-Induced Mood Effects: Decoupling Arousal from Valence," *Journal of Advertising*, 31 (4), 15–26.

- Shimp, Terence A., Eva M. Hyatt, and David J. Snyder (1991), "A Critical Appraisal of Demand Artifacts in Consumer Research," *Journal of Consumer Research*, 18, 273–83.
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22 (11), 1359–66.
- Stanovich, Keith E. (2010), *How to Think Straight about Psychology*, Boston, MA: Pearson Allyn and Bacon.
- Stephens, Nancy (1982), "The Effectiveness of Time Compressed Television Advertisements with Older Adults," *Journal of Advertising*, 11 (4), 48–55, 76.
- Vargas, Patrick T., William von Hippel, and Richard E. Petty (2004), "Using Partially Structured Attitude Measures to Enhance the Attitude–Behavior Relationship," *Personality and Social Psychology Bulletin*, 30 (2), 197–211.
- Ward, Scott, Daniel B. Wackman, and Ellen Wartella (1977), *How Children Learn to Buy: The Development of Consumer Information-Processing Skills*, Beverly Hills, CA: Sage.
- Yoo, Jinnie Jinyoung, and Wei-Na Lee (2016), "Calling It Out: The Impact of National Identity on Consumer Response to Ads with a Patriotic Theme," *Journal of Advertising*, 45 (2), 244–55.
- Yoon, Hye Jin (2016), "Comedic Violence in Advertising: The Role of Normative Beliefs and Intensity of Violence," *International Journal of Advertising*, 35 (3), 519–39.
- Zaichkowsky, Judith L. (1986), "Conceptualizing Involvement," *Journal of Advertising*, 15 (2), 4–14.
- Zhao, Guangzhi, Darrel D. Muehling, and Ioannis Kareklas (2014), "Remembering the Good Old Days: The Moderating Role of Consumer Affective State on the Effectiveness of Nostalgic Advertising," *Journal of Advertising*, 43 (3), 244–55.