1. Laboratory experimentation in marketing

Angela Y. Lee and Alice M. Tybout

Marketing academics, managers, public policy makers, and litigators often ponder questions that involve relationships between alternative treatments or strategies and people's responses. For example, an academic may want to test predictions about how individuals' thinking style may influence perceptions of brand extensions. Or, a brand manager may want to know whether an advertisement highlighting a brand's features is more effective than one highlighting its emotional benefits in generating positive attitudes and intentions to purchase among consumers. A public policy maker may wonder whether a communication using an authority figure or one using "the person next door" will result in a higher percentage of people getting tested for colon cancer. And a litigator contesting patent infringement may seek to establish the extent of lost sales caused by a competitor incorporating a patented design into its products.

A variety of research approaches, including examination of historical data, qualitative research, and consumer surveys, may shed some light on these questions. However, only experiments afford strong causal inferences about such relationships. Although experiments conducted in the field often capture the richness of some real-world situations of interest, experiments conducted in the laboratory can provide a much more rigorous test of a causal relationship and often do so in a manner that contains costs, saves time, and minimizes the risks of competitor response or consumer backlash.

Consider McDonald's, which, like many large companies, has been a frequent target for rumors and myths that can negatively impact sales. A well-known case was the rumor that McDonald's used red worm meat in its hamburgers (Greene, 1978). The company launched heavy TV and print campaigns to counter this false information by using highly credible spokespersons and referencing objective data to debunk the rumor. Although such a response seems intuitively reasonable and is consistent with some basic notions of persuasion, it is not without risk. Theories of information processing suggest pathways by which a direct refutation could be ineffective and may even backfire. For example, if the rumor is deemed to be implausible or not credible, then its refutation could have the undesirable effect of prompting rehearsal of the rumor, thus reinforcing rather than weakening it. Following this line of reasoning, Tybout, Calder,

and Sternthal (1981) conducted a laboratory experiment to examine the effectiveness of three different strategies—the direct refutation message strategy that McDonald's employed, a reframing message strategy that weakened the connection between McDonald's and worms while also suggesting some favorable associations to worms, and a retrieval message strategy that required people to activate prior mental associations toward McDonald's that were unrelated to the rumor. They documented the negative impact of a rumor that McDonald's hamburgers were made with worm meat and the ineffectiveness of the direct refutation strategy McDonald's employed. Further, they demonstrated that the reframing and retrieval strategies that were grounded in information processing theories were effective in countering the negative effect of the rumor on people's attitudes toward McDonald's. Not only did their experiment establish a clear causal relationship between the various damage-control strategies and consumers' attitudes toward McDonald's, it did so in a controlled setting that reduced monetary costs and minimized the potential for adverse publicity or competitive interference that might have occurred had the research been conducted in the field.

THE NATURE OF EXPERIMENTS

What is an experiment? At the most basic level, an experiment is a study in which participants are randomly assigned by the researcher to be exposed to different levels of one or more variables (i.e., independent variables), and the subsequent effect of this exposure on one or more outcome variables (i.e., dependent variables) is observed. Thus, an experiment requires that the researcher identify independent and dependent variables that are of interest for theoretical or practical purposes and seeks to determine whether and how these variables are causally linked.

Why do researchers choose to conduct experiments? Experiments are the best method for establishing a causal relationship between independent and dependent variables because the researcher controls participants' exposure to the independent variable(s), thereby insuring that three conditions required to draw a strong conclusion about causality are met. First, there must be covariation such that changes in the independent variable are associated with changes in the dependent variable. Second, the change in the independent variable or cause must precede the change in the dependent variable or effect in time, a condition referred to as temporal precedence. Finally, no variable other than the independent variable should provide a plausible account for the effect on the dependent variable.

In practice, causal relationships are often posited on the basis of covariation observed in historical data, survey responses, or qualitative research. For example, a manager may examine sales records over time and note that sales declined following price increases. Or, a writer may seize on an association between the level of education of a company's marketing staff and its market share performance, as was done in an *Advertising Age* article announcing, "marketers from companies with significant market-share gains are far less likely to have M.B.A.s than those from companies posting significant share losses" (Neff, 2006). Should the conclusion be that price increases cause sales declines and an MBA education leads to poorer market share performance? Of course not! Although fundamental principles of economics might tempt the manager to conclude that, indeed, raising prices reduces sales, alternative explanations are plausible. Perhaps competitors dropped their prices at the same time the company raised its prices or maybe demand for the company's product varies throughout the year and the price increases happened to coincide with seasonal downturns in demand. Likewise, there are undoubtedly numerous differences between firms that gain versus lose market share other than whether they employ M.B.A.s to manage the marketing function. The share-gaining and share-losing firms may vary in terms of size, industry, geographic location, etc., and these factors could plausibly affect the intensity of competition, as well as many other factors that influence market share. In fact, the causal relationship could be in the opposite direction—low performing firms might be more motivated to hire M.B.A.s than high performing firms. In many situations, managers inferring causality from correlation might seek additional data to rule out alternative explanations, but the alternatives considered are limited to those they can imagine and the *possibility* of additional rivals not addressed by the data always remains.

Ruling out rival explanations is not the only challenge when historical data serve as the basis for causal inferences. It may also be difficult to establish temporal precedence because the determination of the start date of observations is necessarily arbitrary. For example, although most people would expect advertising to influence sales and hence would gauge the effectiveness of advertising by examining sales as a function of advertising expenditure in the same and/or previous period, this approach may distort the true effect of advertising if the firm's budgeting strategy is to spend a certain percentage of last period's sales on advertising. Thus, conducting an experiment in which participants are *randomly assigned* to treatments and the independent variables of interest are systematically manipulated is the best way to establish causality.

Returning to the McDonald's worm rumor study, participants were recruited to come to a lab setting where, under the guise of evaluating a

television program, the rumor was introduced in the treatment condition but not in the control condition. Those who heard the rumor were randomly assigned either to hear a direct refutation of the rumor, a message designed to weaken the association between McDonald's and the rumor, or an assertion that activated associations to the McDonald's brand that were unrelated to the rumor. Their attitudes toward McDonald's were then assessed. Thus, the conditions for establishing causality were met: first, the independent variable (strategy to counter the rumor) was varied before the dependent variable (attitude toward McDonald's) was measured, and a statistically significant covariation between the independent and dependent variables was observed. Further, because participants were randomly assigned to the different treatments or levels of the independent variable, the groups exposed to each treatment were presumably equivalent in the aggregate a priori (i.e., any differences between and within the groups such as age, gender, education level, liking for McDonald's, etc. would not influence the dependent variable systematically). As a result, the sole difference between the groups was the treatment to which they were exposed, making the treatment the only plausible cause for any differences in the dependent variable—attitude toward McDonald's.

Suppose McDonald's management relied on historical sales data to make inferences about the impact of the worm rumor and the effectiveness of the company's refutation strategy. If the data showed a decline in sales following circulation of the worm rumor, and that sales rebounded several months after the company aggressively refuted the rumor, management might conclude that the rumor *caused* a downturn in sales, and further infer that refutation was an *effective strategy* for combatting the negative effect of the rumor on sales.

Tybout et al.'s laboratory experiment suggests that the first, but not the second, inference is warranted. Participants randomly assigned to be exposed to the rumor evaluated McDonald's less favorably than those not exposed to the rumor, ruling out possible rival explanations for the sales decline based on actions by a competitor, or a general downturn in sales for the fast food industry, etc. However, participants randomly assigned to the rumor plus refutation treatment viewed McDonald's just as negatively as those exposed to the rumor but who heard no refutation, suggesting the refutation was *not* effective in countering the rumor's effects and that this strategy should *not* be used in response to future rumors. The rebound of sales might instead have occurred because over time consumers recalled the numerous positive associations they had with McDonald's prior to the rumor, and these associations swamped the impact of the rumor. This interpretation is consistent with the strategies that were found to be effective in the laboratory experiment and suggests that strategies focused on

reducing the connection between the company and the rumor might be an effective strategy in response to any future rumors.

In summary, historical, survey, and qualitative data are excellent sources for hypotheses about relationships between variables, but they are inadequate to support a strong causal inference. In situations where it is important to establish causality, an experiment should be conducted.

CHOOSING BETWEEN LABORATORY AND FIELD EXPERIMENTS

The distinction between laboratory and field experiments is the setting in which the research is conducted. Laboratory experiments occur in settings created by the researcher for the explicit purpose of testing one or more hypotheses. Volunteers are recruited and come to a designated physical or online location where they typically receive some form of compensation in exchange for reacting to certain stimuli presented by the researcher. Although steps are typically taken to disguise the independent variables that are of interest and the researcher's hypotheses, laboratory experiment participants are well aware that they are participating in research and that their responses may have consequences beyond reflecting their own desires. At the same time, when the experimental design exposes participants to a single treatment, the lack of awareness of other conditions reduces the likelihood of hypothesis-guessing even if the induction is relatively transparent.

By contrast, field experiments occur in natural settings where participants encounter treatments and provide responses in what they believe is the normal course of their everyday life. As a result, field experiments allow the researcher to assess the impact of a treatment on real world behavior and not just antecedents of or surrogates for behavior (e.g., attitudes, intentions). However, although the field experimenter may design different treatments and take pains to administer them following random assignment, she has little control over the natural variation of a myriad of variables that are not of particular interest, and the presence of which may make it difficult to pinpoint the relationship of interest even though it exists. Moreover, because participants in field experiments are unaware of their role, ethical issues may arise if the research comes to light at a later point in time. Such was the case when Facebook systematically varied the favorableness of stories in 700,000 users' newsfeeds in order to determine the effect of these stories on users' emotions as reflected in their own postings; or when OKCupid management randomly suggested bad matches to its users in a purported effort to test the validity of its date-matching algorithm.

Whether an experiment is better conducted in the laboratory or in the field depends on how the research findings will be used, as well as the practical concerns mentioned earlier. An experiment may be conducted with one of several goals in mind: (1) testing a theory, (2) testing a theory-based intervention, and (3) establishing a phenomenon or effect and estimating the magnitude of the effect.

Testing a Theory

In a *theory-testing* experiment, the goal is to examine predictions derived from an articulated theory in order to draw conclusions about its merits. The independent and dependent variables are chosen to test the relationships between abstract constructs posited by the theory. The interest lies not in the variables per se, but in the relationships between the theoretical constructs that the variables are assumed to represent. Accordingly, the focus is not on generalizing the magnitude of the specific outcomes observed in the experiment; rather, inferences are made about whether the outcomes are best explained by the theory. If the theory is supported, it may then be applied to situations within a set of relevant domains (see Calder, Philips, and Tybout 1981; Lynch, Alba, Krishna, Morwitz, and Gurhan-Canli 2012 for more detailed discussions).

In order to provide a strong test of a theory, the researcher strives to control extraneous factors that might obscure the relationship between the independent and dependent variables if one actually exists. Failing to detect a causal relationship that exists between the variables is commonly referred to as a Type II error, which is closely related to how much statistical power is afforded given the size of the sample (see later discussion on power). If participants are very heterogeneous, or if variables unrelated to the relationship of interest vary dramatically in the natural environment, the chance of detecting the relationship of interest may be significantly reduced. For this reason, laboratory experiments, which enable the recruitment of a relatively homogeneous sample of participants and afford the researcher control over many variables that are not of theoretical interest, are typically preferred to field experiments when the goal is to test theory.

To illustrate theory-testing laboratory experiments, let's consider the work of Aaker and Lee (2001), which tested hypotheses grounded in regulatory focus theory. Regulatory focus theory distinguishes between two types of goals—promotion goals that involve the pursuit of growth and accomplishment, and prevention goals that involve the pursuit of safety and security. The authors proposed that individuals' view of the self, which may be either independent or interdependent, would moder-

ate whether a message framed in terms of a promotion or a prevention goal would be more persuasive. Specifically, they hypothesized that a promotion goal would be more compatible with an independent self-view, whereas a prevention goal would be more compatible with an interdependent self-view; and that compatibility between self-view and goal would lead to greater persuasion.

Aaker and Lee (2001, exp. 1) tested their hypothesis using a laboratory experiment. Type II error was reduced by using a homogeneous sample—college students from a single university. Participants were randomly assigned to view one of four versions of a fictional website for Welch's Grape Juice that the researchers had constructed to manipulate the two independent variables, self-view and goal type, while holding other features of the website constant. After viewing the website, participants responded to a standard set of questions measuring their attitudes toward and interest in the product. The findings were consistent with the regulatory focus-based hypothesis and no alternative interpretation was apparent. So this research is viewed as supporting and refining regulatory focus theory. The researchers had no interest in the particular sample of participants or in Welch's Grape Juice per se, nor did they attempt to generalize the specific effects (i.e., evaluations of the website) to other samples and stimuli. From the standpoint of their goal of testing a hypothesis grounded in regulatory focus theory, some other homogeneous sample and website or even a print ad for a different brand in a different category could provide an equally rigorous test.

Testing an Intervention

The value of theory ultimately lies in its application to real world situations in the form of theory-based interventions. Researchers may pilot test these interventions prior to implementing them on a grand scale. In an *intervention-testing experiment*, the focus is on the treatments and outcomes rather than on the abstract theory that led to the selection of these variables. The goal is to see whether an intervention or treatment has the desired effect and, if multiple interventions are under consideration, to gauge their relative effectiveness. Rather than striving to create interventions that vary along a single dimension and controlling for factors unaddressed by the theory (as would be the goal in a theory test), researchers often design interventions that operationalize the theoretical constructs in multiple ways so as to maximize the likelihood that the intervention will have the desired impact and relax control over factors that lie outside the theory to better mimic the natural environment to which the results will be generalized.

An intervention-testing experiment may be conducted in either a laboratory or a field setting. The desire to obtain results that generalize to a natural setting would seem to favor conducting intervention tests in the field where the implementation of the intervention and contextual factors cannot be tightly controlled and individuals are unaware of their role as participants. However, testing an intervention in the field can be expensive and time-consuming because it may necessitate implementation on a large scale, and may require the cooperation of a variety of parties whose interests are not readily aligned. Further, companies that operate in a competitive environment may fear that conducting a field experiment could tip their hand to competitors, perhaps allowing them to take actions that distort the test results and even rush a similar competitive product to market. In addition, conducting an intervention test in the field where individuals are unwitting participants can raise ethical concerns and create backlash, as occurred in the case of field experiments conducted by Facebook and OKCupid mentioned earlier. As a result, a researcher may elect to conduct an intervention test in the laboratory. The McDonald's worm rumor study is one such example (Tybout et al. 1981). The researchers drew on theories of information processing to design potential interventions and introduced them in a setting that mimicked one where people might encounter the rumor and McDonald's response to it.

Work by Tal and Wansink (2015) illustrates the use of both laboratory and field experiments to test interventions. These authors drew upon theory about the mental activation of concepts in memory to design interventions that encouraged consumers to make healthy food purchases. Their interventions involved priming either healthy or unhealthy food choices through asking participants to taste (or imagine tasting) food samples (e.g., apple or cookie) and then observing choices they made either on a virtual (laboratory) or actual (field) shopping trip. In all experiments, consumers who were primed to think about healthy choices chose more fruits and vegetables than those primed to think about unhealthy foods, leading the authors to recommend consumers having a small healthy snack before shopping, or grocers offering healthy snack samples in store to promote healthy living.

Establishing a Phenomenon and its Magnitude

Although the desire to test or apply theory is a common motivation for laboratory experiments, researchers may conduct such experiments with the goal of *establishing a phenomenon or the magnitude of an effect* in the absence of a well-articulated, abstract theory. For example, a manager may have an intuition based on sales data across different retail outlets

that sales of a product are tied to its placement within a grocery store such that sales are greater when the product is displayed next to complementary categories rather than potentially competing ones (e.g., peanut butter shelved next to jams and preserves rather than next to soy nut butter). A litigator may need to estimate sales that were potentially lost due to a competitor's infringement on a patent by isolating the effect of specific product features on consumer preferences. Or a charity may desire to select the most effective appeal from several executions for generating donations. In these situations, a field experiment has some obvious advantages. Nevertheless, a laboratory experiment may be the better choice due to monetary and time constraints.

In summary, if the primary goal is to establish a clear causal linkage (versus estimating the magnitude of the relationship in natural settings), a laboratory experiment is preferred. A laboratory experiment may also be preferred for a variety of practical reasons detailed earlier. An important additional advantage of conducting an experiment in the laboratory is the opportunity to solicit participants' responses to other questions that may further shed light on the causal relationship. Information such as age, gender, income, education, past experiences, and their thoughts and emotions while being exposed to the treatment may also be useful in identifying why the effect occurs, when it may dissipate or accentuate, and what kinds of intervention may be useful to enhance or suppress the effect.

DESIGNING A LAB EXPERIMENT

When designing a laboratory experiment, researchers must make a variety of decisions including determining the number of treatments, the manner in which these treatments will be administered, the measures that will be taken to assess the effect of these treatments, how participants will be chosen, and how many participants will be necessary to achieve a reliable inference. Key considerations in making these decisions are discussed below.

Choosing a Passive vs. Active Control Treatment

All experiments have the following elements: independent variables (operationalized by exposure to treatments, denoted by X) and dependent variables (reflecting the observed effect, denoted by O). The simplest design has one independent variable (sometimes referred to as factor) with two levels of treatment, with one of the levels serving as the control

condition. And participants are randomly assigned to each of the treatment conditions.

Experimental Group (EG)	X	O_1
Control Group (CG)		O_2

Participants in the control condition may receive no treatment (i.e., passive control), or they may be exposed to an alternative treatment (i.e., active control). The *no treatment control* option is often included to provide a natural baseline condition to capture the situation where participants behave as they would in the absence of any treatment; although it should be recognized that the mere awareness of participating in research may constitute a treatment of sorts. A no treatment control may be of particular interest when one is considering an intervention and a realistic alternative is to do nothing. If the intervention does not perform substantially better than the no treatment control, it will be difficult to justify allocating any significant time or monetary resources to the intervention.

When the objective of the experiment is to compare the effects of different treatments (e.g., two different versions of an advertisement), the design necessarily involves two *alternative treatments*. An alternative treatment may also be used to achieve a tighter control of the experiment even when the objective is not to test different treatments.

EG 1	X_1	O_1
EG 2	X_2	O_2

For example, a researcher interested in the influence of positive mood on brand choice may prefer to contrast the effects of a positive mood induction (e.g., asking participants to write about a happy event) with a neutral mood induction (e.g., asking participants to write about their most recent trip to the grocery store), rather than a no mood induction. In the absence of any mood induction, participants might arrive at the laboratory varying considerably in their mood based on factors unrelated to the experiment. In general, it is more difficult to detect an effect with a passive, no treatment control than with an active, alternative treatment control.

To examine the effect of salient healthy food choices, Tal and Wansink (2015; exp. 3) conducted a laboratory experiment that included both a passive and an active control group. In the experiment, participants were randomly assigned to one of three treatments: one group consumed a sample of chocolate milk labeled as healthy and wholesome (healthy prime treatment), a second group consumed the same chocolate milk

but labeled as rich and indulgent (unhealthy prime treatment/active control), and a third group received no prime (passive control). The dependent measure was the degree to which participants made healthy food selections in a subsequent shopping trip at an online grocery. The passive control provided a baseline measure of participants' preference for healthy items in the absence of any prime, whereas the active control enabled the researchers to control for the effect of the product used in the prime (i.e., chocolate milk), and to determine the effect of the nature of the prime (i.e., healthy vs. indulgent) relative to the baseline. The findings revealed that the healthy prime significantly increased the number of healthy food choices made relative to both the indulgent prime and the no prime treatments; whereas the number of unhealthy food choices did not vary across treatments. These outcomes suggest that people's food choices are influenced by the salience of healthy options, but not the salience of unhealthy or indulgent options.

While randomly assigning participants to the different conditions is to ensure that any effects observed are due to the difference in treatment, random assignment is sometimes unintentionally violated when researchers assign groups of participants to each of the treatment conditions sequentially over a period of time. This practice is problematic because participants' responses may vary depending on conditions that are not randomly assigned such as the weather, time of day, events reported in the news, and so on. Thus, a better practice is to concurrently assign participants to different treatments each time the experiment is run until the requisite number of participants is achieved.

Between vs. Within-participant Design

When the effect of the treatment is measured by comparing the dependent measures across two different groups as described above, the design is referred to as a *between-participant design*. Alternatively, the researcher may choose a *within-participant design* in which a single group of participants is employed for each level of treatment, and measures of the dependent variable are taken both before and after the treatment.

EG 1 O_1 X_1 O_2

The primary advantage of using a within-participant design is efficiency. By controlling for individual differences, the within-participant design offers the same statistical power in detecting differences using a smaller sample. The disadvantage of the within-participant design is that any effect observed may be open to alternative interpretations. In particular,

the measurement preceding the treatment (O_1) may alert participants to the experimenter's hypothesis or it may simply encourage participants to ruminate about their thoughts and feelings. These factors, alone or in combination with the treatment (X), may account for the change in the dependent measures observed after the treatment (O_2), compromising the ability to make a strong casual inference. However, these concerns can be mitigated if the dependent measures are unobtrusive (e.g., the length of time a participant spends engaging in a task) or are not under participants' conscious control (see discussion of dependent measures later in the chapter).

Single vs. Multiple Factors

When the main objective of the experiment is to compare the effects of different treatments (as in an intervention or effect test), a single factor design with as many levels of treatments as desired may be adequate. However, when the objective of the experiment is to delve into the why or how something happens (as in theory testing), a design involving multiple factors may be needed for at least two reasons.

First, multiple factors may be included for the simple reason that some theories specify moderators or boundary conditions. The simplest multifactor design is a 2 ($X_{A1,}X_{A2}$) × 2 ($X_{B1,}X_{B2}$) design, with participants being randomly assigned to each of the four experimental groups:

	X_{B1}	X_{B2}
X_{A1}	EG 1	EG 2
X_{A2}	EG 3	EG 4

The model of this two-factor design is:

$$y_{ijk} = \mu + \tau_j + \lambda_k + (\tau\lambda)_{jk} + \varepsilon_{ijk}$$

where μ = grand mean, $\boldsymbol{\tau}_j$ is the main effect for the j^{th} level of treatment X_A, $\boldsymbol{\lambda}_k$ is the main effect for the k^{th} level of treatment X_B, and $(\boldsymbol{\tau\lambda})_{jk}$ is the interaction effect for X_{Aj} and X_{Bk}.

As an example, the Aaker and Lee (2001) study that was discussed earlier used a 2 × 2 design to test the hypothesis that individuals' self-view moderates whether a promotion or prevention message frame is more persuasive. The researchers varied the content of a website for Welch's Grape Juice that encouraged participants to adopt one of two self-views (independent or interdependent) and exposed them to a persuasive messages evoking one of two goal orientations (promotion or prevention).

Another reason to include multiple factors is to help rule out alternative explanations. While random assignment to experimental treatments serves to isolate the causal variable, the interpretation of this variable in terms of the construct it represents is not unique. This is because a variable can operationalize multiple constructs (and the reverse is also true—a construct can be operationalized by multiple variables). Thus, simply showing an effect does not allow the researcher to unambiguously establish the proposed relationship. Returning to the Aaker and Lee (2001) study, consider how these researchers represented the construct of self-view in their initial experiment. They did so by varying whether the website for Welch's Grape Juice highlighted benefits of the beverage for oneself (intended to activate an independent self-view) or one's family (intended to activate an interdependent self-view). Although it is reasonable to argue that these treatments represented the construct in the intended manner, they might also have varied participants' involvement in the task, with participants being more involved when the site focused on the benefits of grape juice to themselves rather than to their families. If so, an alternative explanation for the findings could be presented in which involvement and goal focus rather than self-view and goal focus explained the findings. To rule out alternative explanations, multiple variables that might represent the construct could be employed. If the effects of these variables converge, then the plausibility of rival explanations is reduced. This strategy was employed by Aaker and Lee, who used a more elaborate three-factor design in their Experiment 2 to test the relationship between self-view and goal focus, using people's ability to recall the information as the dependent variable. In this study, self-view was varied by priming an independent or interdependent view (as in Experiment 1) as well as by recruiting participants from two different cultures known to be associated with different self-views (American-independent, Chinese-interdependent). They found that American participants as well as those whose independent self-view was activated had better recall of the promotion-framed than the prevention-framed message, whereas Chinese participants as well as those whose interdependent self-view was activated had better recall of the prevention-framed than the promotion-framed message. The convergence of the effects of culture and self-view priming on participants' memory of the message strengthened the theory test that different goal orientations are associated with distinct self-views by limiting the likelihood of a rival explanation of involvement for the results.

In general, adding independent variables to an experiment may increase the rigor of the theory test by ruling out rival interpretations and identifying the specific conditions under which the hypothesized effect occurs. However, this benefit comes with a cost. As the model becomes more

complex, the interpretation of the interaction effects can get progressively more difficult. An alternative to expanding a design to include more factors is to conduct several experiments, each employing a simple 2 × 2 design but differing in context or in the variables used to operationalize the constructs.

Irrespective of the number of factors in the basic design, there may be times when it is desirable to control for the effects of some "nuisance" variables (i.e., factors that lie outside the theory but are likely to introduce systematic variation in participants' responses). For example, if Aaker and Lee (2001) had recruited participants from four different universities or employed websites for not one but four brands, they might wish to control for the idiosyncratic effects of these variables by randomly assigning participants to one of the 16 conditions according to a Latin Square design as illustrated below:

	Brand 1	Brand 2	Brand 3	Brand 4
University 1	A*	B	C	D
University 2	B	C	D	A
University 3	C	D	A	B
University 4	D	A	B	C

Notes:
* A = Independent self-view/Promotion frame.
B = Independent self-view/Prevention frame.
C = Interdependent self-view/Promotion frame.
D = Interdependent self-view/Prevention frame.

This design assumes there is no interaction between the variables of interest (self-view and message frame in this example) and the nuisance variables (participant's university and brand). That is, the effect of self-view × message frame does not vary by university or by brand. And each participant's response is modeled as follows:

$$y_{ijk} = \mu + \rho_i + \beta_j + \tau_k + \lambda_l + (\tau\lambda)_{kl} + \varepsilon_{ijk}$$

where μ = grand mean, ρ_i is the effect of the participant's university *i*, β_j is the effect of brand block *j*, τ_k is the treatment effect of self-view, λ_l is the treatment effect of message frame, $(\tau\lambda)_{kl}$ is the interaction effect for the combination of k^{th} level of self-view and the l^{th} level of message frame.

Full vs. Fractional Factorial Design

When the objective of the research is to test for both main and interaction effects, as is typically the case in theory-testing research, a full factorial design is used where every level of one factor is crossed with all levels of the other factors. This was the case for both of the Aaker and Lee (2001) experiments described above. A full factorial design ensures that all the independent variables in the model, including the interaction terms, are orthogonal to each other so that each of the effects could be estimated independently of all other effects. Sometimes for efficiency it is desirable to use just a subset (i.e., a fraction) of the experimental conditions of a full factorial design, carefully chosen to preserve the orthogonality of the design. With a fractional factorial design, the researcher will be able to estimate the main effects with a much smaller sample, but will not be able to estimate all the interaction effects. One instance of a fractional factorial design is the Latin Square design described earlier. A common use of fractional factorial designs is in conjoint studies (see Chapter 3 on conjoint analysis in this volume).

Another strategy that makes efficient use of participants is to "yoke" additional cells to a simple factorial design. The Tybout et al. (1981) experiment illustrates this strategy. The basic design in this study was a 2 × 2 factorial where the participants were randomly assigned to one of four conditions created by crossing mention of the worm rumor (rumor absent, rumor present) with the inclusion of questions prompting retrieval of prior attitudes toward McDonald's (questions absent, questions present). Two additional treatments were yoked to the condition where the rumor was introduced and the retrieval questions were absent. In the first yoked treatment condition, McDonald's refutation of the rumor was presented. In the second condition, a response designed to weaken the connection between McDonald's and worms while making people's mental associations to worms more positive was presented. The design is depicted below.

	No Rumor	Rumor		
No Retrieval Questions	EG 1	EG 2	EG 5*	EG 6**
Retrieval Questions	EG 3	EG 4		

Notes:
* Rumor, no retrieval questions, McDonald's refutation.
** Rumor, no retrieval questions, a message designed to weaken the connection between McDonald's and worms and making people's associations to worms more positive.

Notice that the yoked treatments could have been included as additional treatments in a fully crossed design by allowing the retrieval questions variable to assume four rather than two levels. Doing so would have required eight cells rather than six cells, while allowing the researchers to examine the effectiveness of dual-approach strategies (e.g., retrieval questions + McDonald's refutation). Yet another design could be a single-factor design with five conditions (EG 1, EG 2, EG 4, EG 5, and EG 6) if the researchers were not interested at all in people's attitudes when prior associations are made salient in the absence of a rumor. The key consideration to bear in mind in design selection is how efficient the design is in serving the objectives of the research.

Choosing Dependent Variables

There are many types of dependent measures that researchers can use to assess the effects of the independent variables in a laboratory experiment. The decision of which measures and how many to include will depend on the goal of the experiment. Theories specify not only outcomes, but also processes by which the outcomes occur. Thus, in testing theories, the researcher may include the outcome measures to capture the proposed effect, such as participants' beliefs about or dispositions toward certain brands or products (i.e., the dependent variable), as well as measures that allow inferences about the process underlying those outcomes (i.e., the mediator variable). These process measures serve to strengthen the test of the theory by allowing the researcher to conduct mediation analyses to uncover the mechanism that drives the proposed effect. By contrast, when conducting an intervention test or seeking to establish an effect, the researcher is primarily interested in whether a desired outcome occurs in response to the treatments, and is less interested in the process that led to that outcome, in which case a smaller set of measures may be included. In the next sections, we describe some of the more commonly used measures in lab experiments.

Self-reported thoughts, mood, beliefs, attitudes, and intentions

Participants may be asked to write down their thoughts in response to different treatments; but more typically, they are asked to report their mood or express their beliefs, attitudes, and intentions using multiple-item rating scales. Some common examples include the Likert scale (strongly disagree–strongly agree), semantic differential scale (e.g., cheap–expensive; very ineffective–very effective), and behavioral intention scale (e.g., definitely would not buy–definitely would buy). Multiple items are often used for each dependent variable so that a more stable indictor of

the underlying construct can be obtained than would occur with a single item. These items are then combined to create an index that serves as the dependent variable in the data analysis.

Choice/behavior

Participants may also be asked to make choices or engage in certain behaviors. For example, they may be sent on an online shopping trip where there are real consequences associated with the choices made (e.g., participants receive these products as compensation for participating in the study). Or participants may be asked to a sample a food product and the amount that they consume is measured as an indicator of their liking. Or, participants may be asked to serve as a spokesperson for a cause and the length and detail of their advocacy may serve as an indicator of the strength of their support for the cause.

Memory and process measures

Participants typically have some control over their responses when they self-report their attitudes and behavioral intentions or make conscious choices. The implicit assumption is that participants have access to their attitudes and feelings, which is not always true. Further, their responses may be subject to the social desirability response bias. The laboratory setting allows the administration of other measures over which participants have less conscious control. These include recall and recognition of stimuli presented in the experiment, reaction times to questions, and physiological measures of attention and arousal such as eye-tracking, galvanic skin response (GSR), electromyogram (EMG), electroencephalogram (EEG), and functional magnetic resonance imaging (fMRI). Inclusion of these measures is particularly useful when the researcher is trying to capture automatic responses. However, physiological measures are expensive to administer on a large scale and their obtrusiveness may be distracting to participants.

Measures of demographic characteristics and individual differences

As noted earlier, when theory testing is the goal, the sample should be relatively homogenous on dimensions not of theoretical interest; whereas when intervention or effects testing is the goal, the sample should reflect the heterogeneity observed in the natural setting to which the researcher hopes to apply the findings. Measures of demographic variables such as age, gender, education, country of origin and income are often included to determine whether the sample has the desired homogeneity/heterogeneity. Demographic variables as well as scales that measure individual differences in personality traits or disposition (e.g., Cacioppo and Petty (1982):

Need for Cognition Scale; Snyder (1974): Self-monitoring Scale) can also be used to operationalize theoretical concepts. This was the case in the Aaker and Lee (2001) experiment discussed earlier where participants' cultural background (American or Chinese) served as one operationalization of self-view. Further, demographic characteristics and individual differences may be used to partition the data post hoc to explore whether the same or different effects are observed in subsets of the sample. Thus, including these measures can be useful in determining the robustness of effects or in exploring potential moderators post hoc.

When multiple measures are included in the design, the researcher must consider the order in which they are presented because there is a risk that initial measures may influence subsequent measures. For example, asking participants to recall information presented in the treatment just before expressing their attitude could alter their attitude by encouraging them to rely on the recalled information that they otherwise may not use. One approach to addressing these concerns is to present the dependent measure of greatest interest first and recognize the potential for order effects on subsequent measures. An alternative strategy is to counterbalance the order of the measures and make order a blocking variable in the design to identify potential biases. In the event an order effect is detected, the researcher may have to consider using dependent variables that are less likely to have an order effect, such as those used to assess nonconscious processes (e.g., response time), or collecting data on these variables using separate experiments.

Selecting a Sample

Historically, participation in a laboratory experiment required people to show up at a physical location. Today, many experiments are still conducted in the physical lab, but a growing number of experiments are conducted online where participants can provide their responses anywhere via a computer or a mobile device.

Online labor markets such as Amazon's Mechanical Turk (AMT), Freelancer, and Guru are now used to recruit research participants. The possibility of conducting research online allows researchers to access a more diverse population other than university students or shoppers intercepted at shopping malls. A recent study comparing samples in political science research found that AMT respondents are more representative of the US population than the convenience samples typically used in in-person experiments, although they are not as representative as, say, a national probability sample (see Berinsky, Huber, and Lenz, 2012). Further, the anonymity afforded by online studies may encourage participants to be

more candid in their responses. However, the biggest disadvantage of using an online labor markets for research participants is the loss of control. When responses are collected online, the researcher has little knowledge of or control over the environment surrounding the participants. Further, the identity of the participant is difficult to verify (Marder, 2015). There is also a growing concern that participants recruited from online pools are savvy, professional survey takers who participate in hundreds of studies per week. As a result, they often become familiar with commonly used experimental manipulations and scales, and the responses they provide may be different from those of a naïve participant that researchers observe in a lab experiment. Thus, researchers using online pools are advised to use novel manipulations to operationalize variables of interest, include different attention checks in the survey to identify those who may be responding to the questions mindlessly without even reading the instructions, and to use a larger sample to reduce the within-cell variance.

Determining Sample Size

How many participants one needs for an experiment depends on several considerations: What is the significance criterion (α)? How much statistical power is desired ($1 - \beta$)? What is the likely effect size (ES)? What test statistic will be used to analyze and interpret the data?

The criterion of statistical significance is the researcher's desire to control for Type I error—the probability of mistakenly "discovering" an effect that does not exist. Typically the maximum risk of committing this error is set to $\alpha = .05$. Another sample size consideration has to do with the power of the experiment. Power refers to the researcher's desire to control for Type II error—the probability of failing to detect an effect that exists. The conventional specification of the Type II error is $\beta = .20$, and the power of the test is $1 - \beta = .80$. The sample size is a function of α, β, and the magnitude of the effect (i.e., ES). Some simple guidelines with illustrative sample sizes are provided by Cohen (1992). For example, to detect a medium difference in means between two groups at $\alpha = .05$ and $\beta = .20$, a sample size of 64 in each condition (i.e., total of 128) is needed; and to detect a small (large) difference, a sample size of 393 (26) per condition is needed.[1]

In the August 2015 issue of the *Journal of Consumer Research*, of the 49 lab experiments reported across the eight empirical papers, the maximum sample size per cell was 189, and the minimum was 9, with a mean of 50 and a median of 42. With most of the effect sizes typically studied in the literature being medium or small, it seemed that many of these studies might be underpowered. However, when researchers use multiple studies

to examine the phenomenon of interest to demonstrate robustness or to identify boundary conditions, the aggregate sample size would likely be adequately powered to detect the effect. Further, there may be additional benefits in running multiple small studies to examine a particular phenomenon over running one large study—it allows the researcher to quantify between-study variation in their quest to test for robustness of the effect across different contexts, thereby allowing for a more efficient estimate of the population average effect size and a better calibration of Type I error (McShane and Böckenholt 2014).

CONCLUDING REMARKS

The focus of this chapter is on when it is appropriate to conduct a laboratory experiment and how to design such an experiment. Experiments are valued for their ability to support strong causal inferences about the relationship between independent and dependent variables. In comparison to field experiments, lab experiments typically afford the researcher greater control over factors that are not of interest and the ability to detect a relationship of interest if it indeed exists. By contrast, field experiments prioritize assessing whether the relationship of interest is powerful enough to emerge despite the "noise" created by the variation in non-focal factors in a natural setting.

To illustrate when a laboratory versus a field setting may be more appropriate for examining a causal relationship, we have described three possible goals that a researcher may have in mind: theory-testing, intervention-testing, and effects-estimation. In theory-testing experiments, the data are valued as evidence for or against some abstract construct relationship; whereas in intervention-testing and effects-estimation experiments, the specific findings are of interest in their own right, either because they indicate how an intervention is likely to perform in a natural setting, or they estimate the magnitude of an effect that is of interest. It is important that this characterization of the three distinct goals not obscure the necessity of some explanation regardless of the researcher's goal. The selection of the independent and dependent variables for investigation presupposes some theoretical explanation, even if the causal model may not be theoretically formalized, as any application of the findings beyond the research setting relies not just on statistical generalization but also the validity of the explanation.

NOTE

1. When comparing between means, Cohen (1988) considered an ES (**d** = $(\mu_1 - \mu_0)/\sigma$) of .20 to be small, **d** = .50 to be medium, and **d** = .80 to be large. When comparing between two proportions (P), he considered an ES (**h** = $\phi_1 - \phi_2$ where $\phi_1 = 2\ arcsin\ (\sqrt{P_k} = 2)$) of .20 to be small, **h** = .50 to be medium and **h** = .80 to be large. And when assessing correlations, **r** = .10 is considered small, **r** = .30 is medium, and **r** = .50 is large.

REFERENCES

Aaker, Jennifer L. and Angela Y. Lee (2001), "'I' Seek Pleasures and 'We' Avoid Pains: The Role of Self-Regulatory Goals in Information Processing and Persuasion," *Journal of Consumer Research*, 28 (June), 33–49.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz (2012), "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Political Analysis*, 20, 351–368.

Cacioppo, John T. and Richard E. Petty (1982), "The Need for Cognition," *Journal of Personality and Social Psychology*, 42(1), 116–131.

Calder, Bobby J., Lynn W. Phillips and Alice M. Tybout (1981), "Designing Research for Application," *Journal of Consumer Research*, 8(September), 197–207.

Cohen, Jacob (1988), *Statistical Power Analysis for the Behavior Sciences.* Hillsdale, NJ: Erlbaum.

Cohen, Jacob (1992), "A Power Primer," *Psychological Bulletin*, 112(1), 155–159.

Greene, Bob (1978), "Worms? McDonald's Isn't Laughing," *Chicago Tribune* (November 20), p. 1, Section 2.

Lynch, John G., Joseph W. Alba, Aradhna Krishna, Vicki G. Morwitz and Zeynep Gurhan-Canli (2012), "Knowledge Creation in Consumer Research: Multiple Routes, Multiple Criteria," *Journal of Consumer Psychology*, 22, 473–485.

Marder, Jenny (2015), "The Internet's Hidden Science Factory," PBS, http://www.pbs.org/newshour/updates/inside-amazons-hidden-science-factory/, February 11 (last accessed October 3, 2017).

McShane, Blakeley and Ulf Böckenholt (2014), "You Cannot Step into the Same River Twice: When Power Analyses are Optimistic," *Psychological Science*, 9(6), 612–625.

Neff, Jack (2006), "Don't Study Too Hard: MBA Marketing," *Advertising Age* (March 20).

Snyder, Mark (1974), "Self-monitoring of Expressive Behavior," *Journal of Personality and Social Psychology*, 30(4), 526–537.

Tal, Aner and Brian Wansink (2015), "An Apple a Day Brings More Apples Your Way: Healthy Samples Prime Healthier Choices," *Psychology & Marketing*, 35(May), online.

Tybout, Alice M., Bobby J. Calder and Brian Sternthal (1981), "Using Information Processing Theory to Design Marketing Strategies," *Journal of Marketing Research*, 18(February), 73–79.