Field Experiments

Veronica Valli, Florian Stahl, and Elea McDonnell Feit

Abstract

Digitalization of value chains and company processes offers new opportunities to measure and control a firm's activities and to make a business more efficient by better understanding markets, competitors, and consumers' behaviors. Among other methodologies, field experiments conducted in online and offline environments are rapidly changing the way companies make business decisions. Simple A/B tests as well as more complex multivariate experiments are increasingly employed by managers to inform their marketing decisions.

This chapter explains why field experiments are a reliable way to reveal and to prove that a business action results in a desired outcome and provides guidelines on how to perform such experiments step by step covering issues such as randomization, sample selection, and data analysis. Various practical issues in the design of field experiments are covered with the main focus on causal inference and internal and external validity. We conclude the chapter with a practical case study as well as a brief literature review on recent published articles employing field experiments as a data collection method, providing the reader with a list of examples to consider and to refer to when conducting and designing a field experiment.

Keywords

Field experiment • A/B test • Randomized experiment • Online experiment • Digital experiment • Business optimization • Causal inference • Experimental design • Internal validity • External validity

V. Valli (✉) • F. Stahl
University of Mannheim, Mannheim, Germany
e-mail: veronica.valli@bwl.uni-mannheim.de; florian.stahl@bwl.uni-mannheim.de

E.M. Feit
LeBow College of Business, Drexel University, Philadelphia, PA, USA
e-mail: efeit@drexel.edu

C. Homburg et al. (eds), *Handbook of Market Research*,
DOI 10.1007/978-3-319-05542-8_3-1

Contents

Introduction 2
Motivation 2
Defining a Field Experiment 4
Experimentation: Causal Inference and Generalizability 8
Estimating the Causal Effect of a Treatment 8
Generalizability of Findings and External Validity 13
Sample Size 15
Experimental Design and Multivariate Experiments 17
Examples of Field Experiments 21
Case Studies 21
Conclusions 26
References 27

Introduction

In God we trust, all others must bring data (Edward W. Deming[1]).

Motivation

Digitalization of value chains and company processes offers new opportunities to measure and control a firm's activities and to make a business more efficient by better understanding markets, competitors, and consumers' behaviors. Among others, the advent of two main sets of methodologies is changing the way organizations do business in the current digital age:

1. *Big Data Analytics*: data mining, machine learning, and other statistical techniques allow practitioners to handle and analyze huge sets of data with a reasonable effort.
2. *Business Field Experiments*: studies conducted outside of the lab by means of easy-to-use software allow managers to reliably answer causality questions at reasonable costs. At the same time, field experiments have become a primary method for investigating scientific phenomena and that is why this chapter considers field experiments aimed at testing theories, of the same importance as those aimed at testing tactical business strategies.

With the primary objective of informing marketing decisions, the fundamental value of market research is the collection, analysis, and interpretation of market-related information (Homburg et al. 2013). Depending on the objective, the research design can be exploratory, descriptive, or causal (Aaker et al. 2011). In particular, the

[1]Edward W. Deming was an eminent engineer, statistician, professor, and management consultant for more than half a century. His work on statistical process control and other strategies for data-driven decision making continues to be relavent today.

causal design is the best approach to identify cause-effect relationships between variables based on preformulated hypotheses (Homburg 2015). Especially for practitioners, answers to the question "does A cause B?" are essential to derive managerial implications (Iacobucci and Churchill 2010), and, in such a context, an experiment is the most suitable and most popular method to establish causality (Crook et al. 2009; Homburg et al. 2013). For example, consider a marketer wishing to know the impact that a 20% discount will have on the proportion of customers making a purchase during a holiday sale. In such a case, comparing sales between a group of customers who were randomly chosen to be offered the discount and another group who was randomly assigned to not receive the offer will give a direct estimate of the incremental sales lift of the discount. For this reason, market researchers and other practitioners are increasingly making use of experiments in the field. Similarly academics have turned to field experiments, when once there was little experimentation outside of the lab. Field experiments are not only applied to inform almost every type of marketing decision (promotions, communications, visual designs, pricing, optimization of digital services, etc.) but also in disparate areas including business organization, product development, health care, human resource management, politics, and so on. As software tools and expertise grow, there are more and more A/B testing case studies showing that the practice of testing is becoming increasingly popular with small- and medium-sized businesses as well as with larger ones (see "► A/B Testing Case Studies" on Optimizely.com for many examples of online field experiments or Applied Predictive Technologies Case Studies on www.predictivetechnologies.com for examples of offline field experiments).

The first field experiments in business practice date back to the first half of the 1900s when experiments revolutionized agriculture and created massive gains in farm productivity. Toward the end of that century, experiments became popular in manufacturing to improve production and quality. At their early stages, especially in firms that focused on product design and engineering, experiments were tremendously costly and often involved the destruction of expensive prototypes, such as in automotive crash testing. Nowadays, the digitalization of value chains has created a data-rich environment that offers both new challenges and new opportunities to managers, policy makers, and researchers, as also recognized in the recent (2014–2016) research priorities of the Marketing Science Institute (MSI). In such an environment, it is possible to measure market response at a much faster speed, allowing managers to track key economic parameters. These tracking skills allow companies to develop more effective business strategies to increase customer retention and loyalty or spending on products and/or services. This increased digitalization has also turned experiments into an economically feasible way to improve marketing decisions. Many marketers are embracing a *test and learn* philosophy with the aid of several platforms, such as Optimizely, Adobe Target, Applied Predictive Technologies (APT), Visual Website Optimizer (VWO), Oracle Maxymiser, and Google Content Experiments, providing easy-to-use software to perform rigorous field experiments in the online and offline environments.

The primary scope of this chapter is to provide an answer to those readers who may be asking themselves: "why should I consider setting-up a field experiment to answer my research or business question?"

As a first answer, bear in mind the following hallmarks of well-designed field experiments:

- Field experiments are one of the most reliable ways to test a theory or to prove that a business action results in a desired outcome.
- Findings from field experiments have direct implications for business operations. In the language of experimentation, we say that they generalize well and have high external validity. On the other hand, lab experiments are acknowledged to have higher internal validity.
- Field experiments are easy to explain to business leaders and policy makers.

Throughout the following pages, we are going to explain each of the aforementioned points in depth advocating a major focus on business-related field experiments and online experiments (A/B tests).

Defining a Field Experiment

Field experimentation represents the conjunction of two methodological strategies: *experimentation* and *fieldwork*.

Defining an Experiment

Experimentation is a form of investigation in which units of observation are randomly assigned to treatment groups. Ex ante randomization ensures that the experimental groups have the same expected outcomes, which is fundamental to achieve an unbiased estimate of the causal effect of the treatment. Experimentation stands opposite to *observational investigations*, in which researchers attempt to draw inference from naturally occurring variations, as opposed to variations generated through random assignment (Gerber and Green 2008). However, some authors (e.g., Teele 2014) prefer to not exclude nonrandomized studies from the group of experiments, while others refer to studies without randomization as quasi-experiments (cf. Campbell and Stanley 1963).

An experiment involves the manipulation of the *independent* (or *explanatory*) variables in a systematic way which is then followed by the observation and measurement of the effect on the *dependent* (or *response*) variable, while any other variables that might affect the treatment are controlled or randomized over (Aaker et al. 2011; Iacobucci and Churchill 2010). For instance, in testing the impact of a 20% off promotion on sales, the researcher manipulates the independent variable of promotion between the two levels of 20% and zero and measures customer purchases as the response variable.

From the perspective of Dunning (2012), true experiments (either in the lab or in the field) show three identifiable aspects:

1. The responses of experimental subjects assigned to receive one treatment are compared to the responses of subjects assigned to another treatment (often a control group which receives some type of baseline treatment that is essentially *no treatment* or the *state-of-the-art* condition). In the case of multivariate experiments, there are several treatment groups, which are all compared among each other.
2. The assignment of subjects to each group is done through a randomization device, such as a coin flip, a dice roll, or a digital algorithm.
3. The manipulation of the treatment is under the control of an experimental researcher.

Some *observational studies* share attribute number 1 of true experiments, in that treatment conditions' outcomes are compared. However, they do not share attributes number 2 and 3 as there is no randomization of treatment assignment and there is no treatment manipulation. On the other side, *natural experiments* share attribute 1 and partially attribute 2 since assignment is random or as-if random. However, in such cases, data comes from naturally occurring phenomena, and therefore the manipulation of treatment variables is not generally under the researcher's control. Natural experiments consider the treatment itself as an experiment and employ naturally occurring variations as a proxy for random assignment. In particular, the treatment is not assigned by a researcher but by some rule-based process that can be mathematically modeled (Teele 2014). Without it, other *confounder* variables could easily explain ex post differences between observed units (Dunning 2012).

Lab Versus Field Experiments

Depending on the setting employed, one can distinguish between laboratory and field experiments (Homburg 2015). In *laboratory experiments*, participants are tested in an environment which is created by the researcher and which thus differs from reality (Aaker et al. 2011). This unreal environment allows the experimenter to control other potential influences on the response but has the main drawback of making the respondent feel observed, which can lead to several kinds of response bias. In addition, the respondents who are willing to participate in a lab experiment may not represent the target population as a whole, and then findings might not be generalizable.

Outside of the lab environment, it is possible to run *field experiments*, in which the setting is an everyday life situation, often the exact same setting where the findings from the experiment will be deployed (Gerber and Greene 2012). In most field experiments, participants are not even conscious of taking part in an experiment (Aaker et al. 2011; Gneezy 2017) eliminating the risk of incurring a response bias. Just as experiments are designed to test causal claims with minimal reliance on assumptions, experiments conducted in real-world settings are designed to make generalizations less dependent on assumptions (Gerber and Green 2012). Further, especially in digital environments such as websites, adequate sample sizes can be much more easily reached than in offline settings or labs, and randomization over large samples protects against the possibility that a variable other than the treatment

is causing the response. Since the aim of this chapter is to provide a complete overview of the topic, a few issues discussed (e.g., issues related to causality, treatment effects, randomization, sources of bias, etc.) apply to experiments in general and therefore to both field and lab experiments. The reader will excuse the unavoidable overlap of some content with other chapters in this book.

Key Features of Field Experiments

Field experiments, either online or offline, can take many forms, but all have four key features that make them a field experiment: authenticity of treatments, representativeness of participants, real-world context, and relevant outcome measures. Indeed, the degree of fieldness of an experiment can vary dramatically; some field experiments may seem naturalistic on all dimensions, while others may be more dependent on assumptions. In a nutshell, what constitutes a field experiment depends on how the field itself is defined (Gerber and Green 2012). Harrison and List (2004) offer a classification system ranking field experiments depending on their degree of realism. The taxonomy they propose is based on six dimensions: (1) nature of the subject pool, (2) nature of the information that the subjects bring to the task, (3) nature of the commodity, (4) nature of the task, (5) nature of the stakes, and (6) nature of the environment that the subject operates in. Harrison and List (2004) propose the following terminology:

- The *conventional lab experiment* employs a convenient subject pool (typically students[2]), an abstract framing, and an imposed set of rules.
- The *artifactual field experiment* is akin to the lab experiment but involving a nonstandard (i.e., non-students) subject pool. With the term artifactual, the authors want to denote studies with an empirical approach that is artificial or synthetic in certain dimensions.
- The *framed field experiment* is akin to the artifactual field experiment but involving a realistic task and the natural environment of the tested subjects that are conscious of being tested. The term framed denotes the fact that the experiment is organized in the field context of the subjects (e.g., social experiments).
- The *natural field experiment* is akin to the framed field experiment involving the environment where subjects naturally undertake the tasks but with the subjects being unaware of participating in an experiment, that is, either online or offline depending on the nature of the setting under examination. Since participants in this kind of experiments are a representative, randomly chosen, and non-self-selected subset of the treatment population of interest, the causal effect obtained from this type of experiment is the average causal effect for the full population, not for a nonrandom subset that chooses to participate (List 2011).

[2]For an interesting discussion on the choice of participants for an experiment and the questionability of employing students, refer to Koschate-Fisher and Schandelmeier (2014).

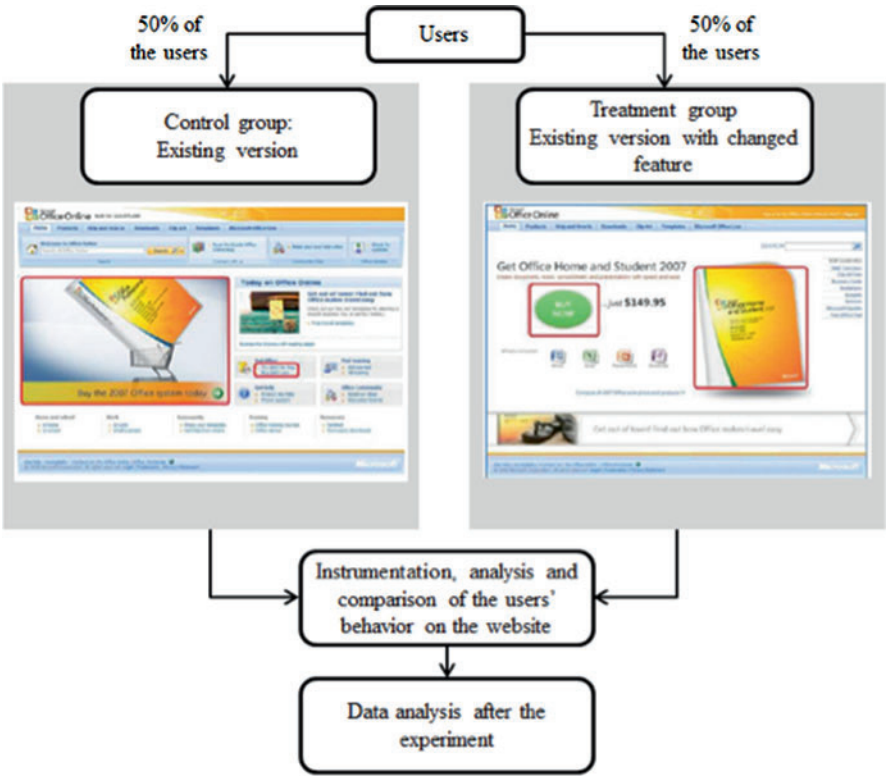

Fig. 1 Example of A/B test on Microsoft Office (Adapted from Crook et al. 2009 and Kohavi et al. 2009)

Online Experiments

Online experiments are a special form of field experiments and their simplest form is commonly referred to as *A/B test*. As shown in Fig. 1, this method involves random assignment of users to two different treatments, typically the current (or A) version and the new (or B) version (Kohavi et al. 2009). In particular, it involves the following steps:

- Randomly divide customers into groups.
- Expose each group to a different treatment.
- Measure one or more selected response variables (also called overall evaluation criteria or key performance indicators, such as conversion rates, click-through rate, revenues, etc.) for both groups.
- Compare groups by mean of data analysis to determine which treatment is better.

Field experiments did not start with digital marketing, and they are certainly not limited to digital marketing, but the digital environment has made testing easier and more popular as a way to inform managers' decisions. Managers are slowly

accepting that carefully observing how customers respond to changes in marketing is more reliable than their experience and intuition.

Experimentation: Causal Inference and Generalizability

Whether it is a simple A/B test to choose the subject line for an email or a complex field experiment to test an economic theory, there are two main issues that the researcher must consider in designing an experiment. The first is whether the experiment has successfully measured the causal effect of the treatment within the context that the experiment is conducted (called *internal validity*). The second is whether the specific findings of the experiment can be generalized to other settings (called *external validity*). In this section, we discuss these two main issues in turn.

Estimating the Causal Effect of a Treatment

The Average Treatment Effect

Field experiments such as A/B tests allow managers to reveal the causal relationship between actions the company might take, such as price promotions (the *cause*), and consumers' purchase decisions (the *effect*). In other words, the goal of a field experiment is to determine whether a particular cause (such as a 20% price promotion) is responsible for an effect (such as a consumer's increased likelihood to purchase a particular product) and to exclude the reverse. Estimating the causal effect of an action has been a golden standard in the social sciences and in economic research for decades, and, as John List (2011) reminds us, economists have long worked on approaches that seek to separate cause and effect in naturally occurring data. For instance, instrumental variable regression aims at isolating cause-effect relationships. Field experiments use randomization as an instrumental variable, which, by construction, is uncorrelated with other variables that might affect the outcome (List 2011). However, there are a few key assumptions that must be met in order for experiments to provide reliable assessments of cause and effect (Gerber and Green 2012, Imbens and Rubin 2015). First, we provide a definition of *causal effect*: a causal effect is the difference between two potential outcomes, one in which a subject receives the treatment and the other in which the subject does not receive the treatment. In formulas:

$$\tau_i \equiv Y_i(1) - Y_i(0)$$

where (τ_i) is the causal effect of the treatment and $Y_i(1)$ is the potential outcome if the ith subject receives the treatment while $Y_i(0)$ is the potential outcome if the ith subject does not receive the treatment. For example, $Y_i(1)$ might be an indicator for whether the customer would make a purchase if she receives the promotion, and $Y_i(0)$ would be an indicator for whether the customer would make a purchase without the promotion.

Of course, it is typically not possible to directly observe both conditions for any given subject, but it is possible to estimate the average treatment effect (ATE) among all subjects, when certain assumptions are met. The ATE is defined as the sum of the subject-level treatment effects, $Y_i(1) - Y_i(0)$ divided by the total number of subjects. In formulas:

$$ATE \equiv \frac{1}{N} \sum_{i=1}^{N} \tau_i$$

The challenge in estimating the ATE is that at a given point in time, subject *i* is either treated or non-treated, and therefore either $Y_i(1)$ or $Y_i(0)$ is observed, but not both. Some statisticians conceptualize this as a missing data problem where either $Y_i(1)$ or $Y_i(0)$ is unobserved for each subject (Imbens and Rubin 2015).

Experiments, both in the lab and in the field, provide unbiased estimates of the ATE when the following assumptions are met (Gerber and Green 2012):

1. *Random assignment*: treatments are allocated such that all units have an equal probability between 0 and 1 of being assigned to the treatment group.
2. *Excludability*: the treatment must be defined clearly so that one can assess whether subjects are exposed to the intended treatment or to something else.
3. *Noninterference*: no matter which subjects the random assignment allocates to treatment or control, a given subject's potential outcomes remain the same.

Let us consider the three assumptions in more depth.

Random assignment is fundamental in experimentation, with roots that go back as far as Neyman (1923) and Fisher (1925). It implies that treatment assignments are statistically independent of the subjects' potential outcomes and addresses the missing data issue that challenges the estimate of the ATE, that is, the issue that at a given point in time, subject *i* is either treated or non-treated and therefore either $Y_i(1)$ or $Y_i(0)$ is observed, but not both. In fact, when treatments are allocated randomly, the treatment group is a random sample of the population in the experiment, and therefore the expected potential outcomes among subjects in the treatment group are identical to the average potential outcomes among the control group. Therefore, in expectation, the treatment group's potential outcomes are the same as the control group. When units are randomly assigned to treatment and control, a comparison of average outcomes in treatment and control groups, the so-called *difference-in-means* estimator, is an unbiased estimator of the ATE. In formulas, the estimator is:

$$\frac{1}{N} \sum_{i \in \text{treated}} Y_i - \frac{1}{M} \sum_{i \in \text{control}} Y_i$$

where *N* is the number of subjects in the treatment group and *M* is the number of subjects in the control group. We can see that the expected value of the estimator is equal to the ATE, meaning it is unbiased:

$$E\left[\frac{1}{N}\sum_{i\in\text{treated}} Y_i - \frac{1}{M}\sum_{i\in\text{control}} Y_i\right] = E\left[\frac{1}{N}\sum_{i\in\text{treated}} Y_i\right]$$
$$-E\left[\frac{1}{M}\sum_{i\in\text{control}} Y_i\right] = E[Y_i(1)] - E[Y_i(0)] = E[\tau_i] = ATE$$

When random assignment is not used, there is always potential for a *selection bias*, where the treatment assignment is systematically related to potential outcomes. For example, if we want to measure the effect of a call from a sales agent and we do not randomize calls between customers, the sales agent may choose to call those customers that she/he feels are most likely to buy. This will produce an upward bias in our estimate of the ATE. The key idea is that randomized assignment allows us to use simple averages of the outcome for the treatment and control group to estimate the average treatment effect.

Excludability refers to the fact that each potential outcome depends *solely* on whether the subject *itself* receives the treatment and not on some other feature of the experiment. Therefore, when conducting an experiment, we must define the treatment and distinguish it from other factors with which it may be correlated. Specifically, we must distinguish between d_i, the treatment, and z_i, a variable that indicates which observations have been allocated to treatment or control. We seek to estimate the effect of d_i, and we assume that the treatment assignment z_i has no effect on the outcomes. In other words, the exclusion restriction refers to the assumption that z_i can be omitted from the potential outcomes for $Y_i(1)$ and $Y_i(0)$, and this restriction fails when random assignment sets in motion causes of Y_i other than the treatment. In real life, and therefore in field experiments in particular, it can become difficult to ensure excludability. Consider, for example, an A/B test investigating the impact of a discount on purchase decisions. If being assigned to receive a discount also means that the customer will get an email and customers in the treatment group do not get an email, then the excludability assumption is not met, and any observed difference between the treatment and control groups may be due to the email and not to the discount. A straightforward example of a research design that attempts to isolate a specific cause is a pharmaceutical trial in which the treatment group is given an experimental pill while the control group is given an esthetically identical sugar pill. The aim of administering a pill to both groups is to isolate the pharmacological effects of the ingredients, holding constant the effect of merely taking some sort of pill (*placebo effect*).

How can we make sure that the excludability assumption is met and that we are able to isolate the specific cause we intend to? Basically, by ensuring uniform handling of treatment and control groups, for instance, with double blindness, neither the subjects nor the researchers charged with measuring outcomes are aware of which treatments the subjects receive, so that they cannot consciously or unconsciously distort the results. Another procedure is parallelism when administrating an experiment: the same procedures should be used for both treatment and control groups, and both groups' outcomes should be gathered at approximately the

same time and under similar conditions (Gerber and Green 2012). In online experiments, meeting excludability assumptions might seem easier; however, consider, for instance, a test of several versions of the same webpage showing prices and promotions for a given brand. Randomization algorithms ensure that different customers shopping from different laptops and IP addresses see different versions. But if, unluckily, two people sitting one next to the other and surfing the same webpage from different terminals but in the same location see different versions (having been assigned to different treatment groups), we might incur in a violation of the exclusion restriction, as recognizing the different versions can confound the causal effect we set out to estimate. In such cases, precise geolocalization and a randomization procedure that considers such geographical information could help solve the problem.

Noninterference refers to the fact that potential outcomes are defined over the set of treatments that the subject *itself* receives, not the treatments assigned to other subjects. This assumption is sometimes called the Stable Unit Treatment Value Assumption (SUTVA). Considering that each observational unit is either treated or not treated, the number of potential outcomes to take into account can quickly increase if we allow the outcome for subject i to depend on the treatment assignment of another subject j. The noninterference assumption cuts through this complexity by assuming that the outcome for i is not affected by the treatment of other subjects (Gerber and Green 2012; Imbens and Rubin 2015). Consider, for instance, when an A/B test is conducted on an e-commerce website offering promotions to a targeted subsample of existing customers and not to some others. Noninterference would assume that purchase decisions of subject i were only affected by his/her personal assignment to treatment or control group. But what if, for instance, two subjects belonging to the same household, say two sisters, are shopping from the same website and one falls into the treatment but the other one falls into the control? Then, we might have violation of noninterference as the treatment received by one sister can affect the other that therefore no longer constitutes an untreated control group. To prevent this from happening, researchers should try to design experiments in ways that minimize interference between units by spreading them out temporally or geographically or to design experiments in ways that allow researchers to detect spillover between units. Instead of treating interference as a nuisance, these more complex experimental designs aim to detect evidence of communication or strategic interaction among units.

Causality and Internal Validity

The previous section described the issues involved in estimating the causal effects as they are typically discussed in economics (Gerber and Green 2012) and statistics (Imbens and Rubin 2015). Psychologists also have a rich tradition of describing problems that can occur in experiments and have coined the term *internal validity* which refers to the extent to which we can say the observed effect in our study was caused by our treatment (Campbell 1957; Campbell and Stanley 1963; Shadish et al. 2002). Many of the ideas in this section are closely related to the previous discussion

of the conditions necessary to estimate the causal average treatment effects, but using a different set of terms. Since both perspectives on experiments are common in marketing, we present both.

To achieve high internal validity, laboratory experiments are generally more suitable. This is because the controlled environment allows for better control of confounders. However, depending on the field considered, the natural environment can be highly controlled as well, especially in digital settings. In general, when considering studies that go beyond the randomized controlled experiment, there are many threats to internal validity, some of which we have discussed previously and most of which apply to both field and lab experiments:

- *Selection bias*: when assignment to treatment is not random and certain types of people are more likely to receive one of the treatments, in other words the experimental groups systematically differ from each other either because of self-selection (e.g., by voluntary choosing whether to receive the treatment) or by incorrect assignment (Campbell 1957; Iacobucci and Churchill 2010; Shadish et al. 2002). For example, when running an offline field experiment to test the effect of marketing actions on purchase intentions, a selection bias could emerge due to self-selection of respondents into treatments. When treatments are not randomly assigned, the subjects or the experimenters may assign certain types of subjects to treatment and other types to the control. For example, if we are studying the effect of receiving emails on customer's purchase rate using observational data collected by the company, we have to consider that customers get to self-select whether to sign up for the mailing list, and so those who sign up may be systematically more likely to purchase than those who do not sign up. This is less likely to happen in online field experiments, as assignment to treatment groups is handled by the computer systems and mostly unnoticed by users who are often completely unaware of being tested.
- *Differential attrition*: when certain types of subjects drop out of one of the treatments. It implies that certain types of participants leave during the run of the experiment or do not take part in the final measurement (Aaker et al. 2011; Shadish et al. 2002), and this attrition is different for the treatment and the control groups. For instance, if you were testing an increase in the frequency of direct marketing, customers who have less affinity for the brand may be more likely to ask to be put on a "do not call" list when they are in the high-frequency condition. These participants would not complete the treatment and so typically would not be counted in the analysis of the response. The direct consequence of differential attrition is that the average of the experimental group might differ if the exited participants were still involved (Iacobucci and Churchill 2010; Shadish et al. 2002).
- *Time effects*: when treatments are administered at two different times, outside events, learning, or other changes are confounded with the treatment (Shadish et al. 2002).
- *Confounding variables*: when other variables are correlated with the treatment and have an effect on the outcome, a cause-effect relationship between the

confounder and the dependent variable can be mistakenly assumed to be a causal effect of the treatment.

- *Noncompliance*: subjects assigned to the experiment do not get the specified treatment. This can happen because of individuals' voluntary decision to use a different treatment than the one they were assigned, because they do not like it or they think another treatment would be better.
- *Diffusion of the treatment across groups*: subjects assigned to one treatment find out about the other treatment.
- *Demand effects*: participants guess the hypothesis of the experiment and try to cooperate by exhibiting behavior that confirms the hypothesis.
- *Experimenter bias*: experimenter makes subjective measurements and inadvertently favors the hypothesis in those measurements. An experimenter bias may exist when the mere presence or interaction with the interviewer has an effect on the respondent's responses. Being interviewed about personal purchase intentions might arouse a sense of self-exposure that could lead to biased responses not reflecting the private true intentions. This is more often the case in face-to-face interviews and is quite unlikely to happen in lab experiments or in online field experiments.
- *Hawthorne effect*: it is also possible that individuals being part of an experiment and being monitored change their behavior due to the attention they are receiving from researchers rather than because of the manipulation of the independent variables. The Hawthorne effect was first described in the 1950s by researcher Henry A. Landsberger during his analysis of experiments conducted during the 1920s and 1930s at the Hawthorne works electric company in Illinois. His findings suggested that the novelty of being research subjects and the increased attention deriving from this could lead to temporary increases in workers' productivity. This is sometimes also referred to as the *John Henry effect* and is closely related to the *placebo effect* in medicine. This issue is easily overcome in many field experiments where subjects are unaware of being a subject in a test but is more likely to happen in lab experiments (Landsberger 1958).
- *Ambiguous temporal precedence*: In some experiments, it can be unclear whether the treatment was administered before or after the effect was measured. For instance, if purchases and promotional emails are tracked at a daily level, it can be difficult to discern if a customer who received an email on a particular day and also made a purchase that same day received the email before she made the purchase. If the treatment does not occur before the outcome is measured, then the causality may be reversed.

Generalizability of Findings and External Validity

Often, we are interested in whether the conclusions of our experiment can be applied to a specific business decision. For instance, if we test a new product display in 30 stores within a chain and find that the new product display increases sales, then we want to know whether this finding will generalize to other stores in the chain or to

other retailers. *External validity* refers to the extent to which the specific findings of the experiment can be generalized to other target populations or other similar situations (Campbell 1957; Shadish et al. 2002). If the study shows high external validity, we can say that the results can be *generalized*. Field experiments are largely acknowledged to better generalize to real situations than lab experiments because of the real setting in which they are deployed, although some have cautioned that field experiments conducted in one setting cannot always be generalized to other settings (Gneezy 2017).

The major threat to external validity is that some idiosyncrasy of the test situation (*context effect*) produced the effect, but the effect goes away in the target business environment. For instance, while an ad may perform well in a copy test where customers are brought into a lab setting and exposed to the ad and then surveyed on their purchase intent, those results may not generalize to ad exposures in the real world, perhaps because people do not pay as much attention to ads in the real world as they do in the lab. Or a finding from a field experiment showing that price promotions increase sales of packaged goods may not extend to a different product category. For those familiar with regression, another way to conceptualize context effects is that there is an interaction between the treatment and some context variable that was held fixed in the experiment, such that the effect of the treatment is different depending on the value of that context variable (Campbell and Stanley 1963).

Another key element in designing an experiment with good external validity is determining which subjects to include in the experiment. Note that the assignment of subjects to treatments is closely related to the internal validity of the test, while the selection of subjects to include in the experiment is closely related to the external validity. The best way to enhance external validity is to test the research hypotheses on the entire population that the researcher hopes to learn about, e.g., all the customers in a CRM system or all the stores in a chain. This approach also maximizes the power of the test to detect differences between treatments, which we will discuss in the next section. Obviously, this is rarely possible outside of some digital marketing contexts either because of the high costs of applying treatments and measuring outcomes and/or the riskiness of the treatment.

To reduce risks and costs, researchers frequently rely on samples of subjects from the target population. Some sampling strategies that are available to use are (from ideal to worst):

- *Simple random sample:* take a random draw from the target population using, for instance, a coin flip or a dice roll. This gives to each subject an identical probability of entering the sample, ensuring that the sample will be representative of the target population.
- *Cluster sample*: when it is easy to measure groups or clusters of subjects, randomly sample from among the clusters.
- *Stratified sample*: use a procedure to make sure that the sample contains different types of subjects.

- *Convenience sample*: sample in some way that is easy for the researcher, e.g., an academic might conduct the experiment with students or a company might conduct the experiment using store locations that are nearby.

For instance, if a publishing company wants to evaluate whether a given promotion strategy works better than another and decides to run a field experiment, they have to consider the target population from which to sample. If their goal is to learn how their current customers respond, they might focus on customers from their current mailing list. However, if they hope to learn about how *potential* customers respond to the promotions, they might choose to sample customers from a larger list of avid readers. In either case, once the target population is identified, the ideal strategy for selecting a group of customers to include in the experiment is to either use all the customers in the target population, assigning some to treatment and some to control, or to select smaller treatment and control groups randomly from the mailing list. The simple random sample ensures that the subjects in the study represent the target population. A convenience sample, by contrast, may not properly represent the target population; for example, students may not behave in the same way as other types of customers. If the company plans to study separate subgroups within the target population, they may find a stratified sample useful for ensuring that there is sufficient sample size within each subgroup. Another potential threat to generalizability is the representativeness of the subjects in the test. A common criticism of experiments conducted with students, for instance, through surveys or lab experiments, is that the results may not reliably extend to the entire population of reference. Similarly, in online experiments the researcher should keep in mind that mostly heavy users of the website or app are more likely to be included in field experiments than light users. Most online tests include in the sample all the visitors in a fixed period, and this group will naturally include more frequent users than infrequent users. To overcome such issues, companies should consider test designs that assign treatments to users (rather than to sessions), track users across visits, and cap the number of times each user is exposed to the treatment.

Sample Size

A key question in designing any experiment is determining how many subjects to include in the test. Sample sizes for an A/B test are typically determined by considering the hypothesis test comparing the two groups. The typical A/B test in marketing estimates the average treatment effects by comparing the proportions of people who respond to two different stimuli. Following the traditional one-tailed test for comparing proportions, we begin with a null hypothesis that the proportion of people who respond will be the same in both groups versus an alternative that the A group responds in greater proportion than the B group:

$$H_0 = \pi_A = \pi_B = \pi$$
$$H_1 = \pi_A - \pi_B = \delta > 0$$

Our goal is to plan the number of subjects to include in the treatment and control groups so that we will be able to correctly retain the null hypothesis if there is no difference between treatments and reject the null if there is a difference of at least δ. In the extreme, if we have no subjects, then we clearly will always retain the null hypothesis no matter what. There are four aspects of the experiment that influence the expected required sample size for an A/B test:

- The expected proportion π
- The expected (minimum) difference between the two groups δ
- The desired confidence 1-α (where α is the significance)
- The desired power 1-β

The *confidence* is the likelihood that you will retain the null hypothesis and decide that there is no difference when there really is no difference. *Power* is the likelihood that you will reject the null and detect a difference when indeed there is a difference of at least δ. Both should be considered carefully in the design of an experiment. Consider, for example, an A/B test designed to determine the effect of an ad on the proportion of people who buy. In this case, we want high confidence to prevent the possibility of concluding that the ad has a positive effect when it, in fact, does not. We also want high power, to prevent concluding that that the ad does not work when, in fact, it does. For a given sample size, power and confidence can be traded off. Lewis and Rao (2015) find that for display advertisements, even A/B tests with very large sample size conducted at a traditional confidence level of 0.95 do not have sufficient power to detect whether an ad has positive ROI. Thus, it is critical to consider power when planning an A/B test.

The sample size for each group in a comparative A/B test can be accurately estimated by (Ledolter and Swersey 2007):

$$N \approx \frac{2\pi(1-\pi)\left[z_{1-\alpha} + z_{1-\beta}\right]^2}{\delta^2}$$

where z_x is the cumulative normal distribution evaluated at x. This can be computed, for example, using the Excel formulas: $z_{1-\alpha} = NORM.S.INV(1-\alpha)$ and $z_{1-\beta} = NORM.S.INV(1-\beta)$.

One can see from this formula that if the researcher wants to detect a small difference, δ, in the response rate between the A and B groups, then a larger sample size is required. Similarly, if the researcher wishes to reduce the chance of an erroneous conclusion (i.e., that there is a difference when there is not or that there is not a difference when there is), then $z_{1-\alpha}$ and $z_{1-\beta}$ will be larger and the required sample sizes will be higher.

Note that this formula depends on the size of the difference that the marketer wishes to detect. In practice, it is very important to consider δ carefully. When a very

large amount of data is available (for instance, from e-commerce websites), generating large datasets and big samples is much easier than few years ago. In such cases, it might happen that very negligible effects become significant (e.g., WTP is $10 in treatment group and $9.99 in control). While this effect is statistically significant, it does not really tell much about our business/research question and may not be useful for making decisions. So, in situations where N is not limited by the budget, it may be sensible to choose a smaller N so that the difference to detect, δ, is a difference that would be meaningful to the business. This is sometimes referred to as aligning practical and statistical significance.

Experimental Design and Multivariate Experiments

Managers frequently want to measure the effect of several different marketing actions (i.e., they are interested in more than one treatment). For instance, a publisher might be interested in assessing how different discount levels perform in combination with different ways of communicating the discount. They might be interested in measuring the effect of two levels of discount (say 5€ and 10€) while at the same time understanding the effect of communicating the price reduction in terms of price discount (e.g., "subscribe for one month and save x €!") or in terms of bonus time (e.g., "subscribe for 1 month and get x weeks free!"). A multivariate experiment can be used to simultaneously measure the effect of the discount level and the message type while also determining if there is any additional effect of combining two treatments together. When the combined effect of two treatments is better than the sum of the individual effects, there is an *interaction* effect. Detecting interactions is the main reason why companies conduct multivariate tests. In addition, multivariate tests can reduce required sample sizes and increase the amount that can be learned in the time frame of a single test.

Before approaching the technicalities of multivariate testing, we define some useful terminology. The *factors* are those variables (continuous or categorical) whose effect we want to study, e.g., ad copy, font, photo, and color in an advertisement or seed type, fertilizer, and amount of water for an agricultural experiment. In the experiment, each factor is tested at multiple *levels*, the different versions we want to test. The simplest A/B test comparing two treatments has 1 factor with two levels.

Multivariate tests are experiments where two or more factors are tested. Multivariate tests should be carried out when the researcher wants to know the relative effects of the different factors or when there might be combinations of levels that perform especially well together. If the effect of the two factors together is more (or less) than the sum of their separate effects, we say the two factors interact with each other. For instance, the text color and the background color of a call-to-action button typically interact: when the colors are the same, customers cannot read the button and do not respond.

For a better understanding of multivariate experiments, consider the following experiment (adapted from Ledolter and Swersey 2007) that was conducted by a credit card company who wanted to increase the response rate, that is, the number of

people who respond to a credit card offer. The marketing team decided to study the effects of interest rates and fees, using the four factors shown in the following table.

	Factor	Level 1 (−)	Level 2 (+)
A	Annual fee	Current	Lower
B	Account-opening fee	No	Yes
C	Initial interest rate	Current	Lower
D	Long-term interest rate	Low	High

We could choose to study these factors with a series of A/B tests. Suppose we all agree that factor A (annual fee) is likely to be most important. Then we can run an A/B test on annual fee, holding the other factors at the control levels. The combination of factors and levels is clearly summarized in the following *design matrix*:

Run	A Annual fee	B Account-opening fee	C Initial interest rate	D Long-term interest rate	Sample
1	–	–	–	–	20,000
2	+	–	–	–	20,000

Suppose our first test found that the lower annual fee increased the response rate. So, we can fix the factor A to "+" and in our next A/B test, we can look at factor B:

Run	A Annual fee	B Account-opening fee	C Initial interest rate	D Long-term interest rate	Sample
3	+	–	–	–	20,000
4	+	+	–	–	20,000

Putting a sequence of these A/B tests together, we might end up with the following runs:

Run	A Annual fee	B Account-opening fee	C Initial interest rate	D Long-term interest rate	Sample
1	–	–	–	–	20,000
2	+	–	–	–	20,000
3	+	–	–	–	20,000
4	+	+	–	–	20,000
5	+	–	–	–	20,000
6	+	–	+	–	20,000
7	+	–	–	–	20,000
8	+	–	–	+	20,000

Looking back at the resulting set of runs, we might notice several serious problems:

- Before we run the first A/B tests, we do not really know which factor is most influential, so it is difficult to know where to start.
- We could be wasting time with the sequential process.
- We are sometimes running the same condition more than once, which is inefficient (runs 2, 3, 5, and 7 are all the same).
- Because we have not tested all combinations of factors, we have little information about the interactions between factors.
- If there are interactions, testing the factors in a different sequence could lead to different conclusions about which combination is best.

To overcome these issues, it is recommended to make use of a proper experimental design (commonly referred to as design of experiment, or DOE). In this example, a better approach creates a single test that includes every possible combination of levels (*full factorial design*) which allows us to see if there are certain combinations of factors which are particularly good and to reduce the sample sizes for each run. The full factorial design matrix, in this case, looks like this:

Run	A Annual fee	B Account-opening fee	C Initial interest rate	D Long-term interest rate	Sample
1	–	–	–	–	7500
2	+	–	–	–	7500
3	–	+	–	–	7500
4	+	+	–	–	7500
5	–	–	+	–	7500
6	+	–	+	–	7500
7	–	+	+	–	7500
8	+	+	+	–	7500
9	–	–	–	+	7500
10	+	–	–	+	7500
11	–	+	–	+	7500
12	+	+	–	+	7500
13	–	–	+	+	7500
14	+	–	+	+	7500
15	–	+	+	+	7500
16	+	+	+	+	7500

Note that the number of possible combinations for a design can be computed by multiplying together the number of levels (2) for each of the four factors ($2 \times 2 \times 2 \times 2 = 2^4 = 16$ combinations). It has become common to describe an experiment with multiple factors using this shorthand. For example, a $2^3 \times 5^1$ full factorial experiment has three factors that have two levels and one factor that has five levels, which is a total of 40 different combinations of the factors.

A full factorial design allows us to estimate the *main effects* of the factors and all *interactions* between factors. The *main effect* of a factor is defined as the change in

the response variable when the level of the factor is changed from low to high and corresponds to the average treatment effect for an A/B test that we discussed in Sect. 2. For a full factorial design, we can compute the main effect, by averaging the response rate across the runs when the level is at the high level and comparing that to the average across the runs at the low level. A two-way *interaction* occurs when the effect of one factor depends on the level of another factor (e.g., does the impact of having an annual fee depend on whether or not there is an account-opening fee?). *Three-* and *four-way interactions* are similar to two-way interactions, but are difficult to think about (e.g., is the effect of C different when both A and B are at their high levels?). Luckily, those higher-order interactions are usually negligible in most business settings. To estimate main effects and interactions for multivariate experiments, most researchers use *regression analysis*, fitting a model that relates the outcome measure to the various factors. If the subjects in a multivariate test are assigned to conditions randomly, the estimates of main effects and interactions that we get from this regression represent the *causal* effect of those treatments, just as in single-factor experiments.

In the example above, we show a full factorial test, where all the possible combinations of factors are tested. However, as the number of factors increases, the number of combinations increases rapidly. Therefore, researchers who use multivariate tests frequently spend a lot of time thinking about which combinations of factors they should include in their experiment and which they can leave out. One approach is *fractional factorial* design, which reduces the number of combinations to a half or a quarter of the possible combinations, by eliminating the possibility of estimating high-order (three-way and higher) interactions. A newer approach for determining which combinations of factors to include in a multivariate test is *optimal design*, which characterizes how much we learn from an experiment by considering how precisely we will be able to estimate the parameters of our regression model. Optimal designs choose the design matrix so as to get the best possible standard errors and covariance matrix for the parameter estimates. (See Goos and Jones 2011 for more detail.) Optimal design typically requires specialized software (e.g., JMP from SAS or the AlgDesign package in R) where the user inputs the factors and levels and the software finds the best combination of factors to test.

An important feature of good multivariate experimental designs is *orthogonality*. When two variables are orthogonal in an experiment, it means that the various combinations of the two factors occur exactly the same number of times. A nice property of orthogonal design is that the estimate of the effect of one factor will not depend on whether or not the other factor is controlled for in the regression. When the two factors are always set at the same level (e.g., the account opening fee is always paired with the annual fee), it is impossible to estimate separate effects for each factor, and this is called a *confound* in the multivariate design, which is the opposite of orthogonality. Full and fractional factorial designs maintain orthogonality, while optimal designs are not necessarily orthogonal, but are usually nearly orthogonal.

One common application of multivariate testing in marketing is in testing various features of direct mail offers: from the color of the envelope to the celebrity

endorser's appeal. In this type of experiment, the direct marketer typically sends out a number of different direct marketing offers with varying levels of the features and then measures the number of customers who respond. In this context, additional cost is incurred for each different version of the mailing, and so fractional factorial and optimal design approaches, which reduce the number of required combinations, are valuable. Applying an optimal design or an orthogonal, fractional factorial design instead of a one-factor-at-a-time method increases the efficiency at evaluating the effects and possible interactions of several factors (independent variables).

Another important application of multivariate experimental design is *conjoint analysis*. In conjoint analysis, customers are asked to evaluate or to choose from a set of hypothetical products, where the products vary along a set of features. These product features become the factors in a multivariate experimental design. A common approach to creating the questions to include in a conjoint survey is to use optimal design (Sándor and Wedel 2001).

Examples of Field Experiments

Case Studies

Field Experiments in Business

Field experiments are rapidly becoming an important part of business practice, and many marketing-oriented firms now employ a testing manager, who is responsible for designing, executing, and reporting on field experiments to answer important questions. These testing managers often specialize in a particular part of the business or communication channel. For instance, one might find different specialists in website testing, email testing, and direct marketing experiments, all within the same company. Regardless of the specific platform, the goal of these testing managers is to find treatments to test, to determine how to measure the response to the treatments, to ensure that the test is designed so that it can be interpreted causally, and to analyze and report on the results. In the next subsection, we describe the testing program employed by the donation platform for the 2012 US presidential campaign for Barack Obama.

A major focus for the 2012 US presidential campaigns was fundraising. Several changes in regulation had made donations to political campaigns more important than ever, and so there was a major focus on the web platform where potential donors were encouraged to make small- and medium-sized donations. In their ongoing efforts to improve the platform, the team conducted more than 240 A/B tests over 6 months to determine which marketing messages worked best (Rush 2012a).

An important consideration for any testing team is deciding which features of the website platform to test. The ultimate determination of which features are worth testing should depend on the potential returns the firm can gain by acting on the findings of the test. The potential returns depend both on how much better the new treatments perform (which is of course unknown before the test) and how many customers will be affected by the treatment. Consequently, most testing teams

choose to test features of their marketing that are seen by many customers and that they believe have a large potential to increase sales or other desired outcomes.

The team managing the donation platform for the Obama campaign tested several areas of the website including imagery, copy, and the donation process. Figure 2 shows an example of an image test that was used on the splash page, where potential donors arrived after clicking on a link describing a special campaign where donors could win a "▶ Dinner with Barack" (adapted from Rush 2012b). The objective of the test was to learn whether the focused shot showing the candidate smiling (which they labeled as *control*) would perform better than the wide shot showing several attendees at a previous event chatting with the candidate and his wife (which they labeled as *variation*). Previous tests had shown that large images of the smiling candidate increased the donation rate, so the team hypothesized that the control image would perform better. The images were assigned randomly in real time to all visitors who clicked on a link to the splash page. The team used the Optimizely web-testing platform, which, like other web-testing platforms, handles the random assignment of treatments automatically and integrates with the web analytics platform to measure the response. The team assessed the performance of the two images, by comparing the percentage of people who made donations in the control group relative to the variation group. The team found that the wider shot showing previous guests at the table with the candidate resulted in a 19% increase in donations. Based on this finding, they quickly decided to change the splash page to the variation image for the remaining duration of the campaign.

Figure 3 shows another example of a test described by Rush (2012b) that involved website copy. The website had a feature that invited donors to store their payment information so that they could make donations in the future with one click. This was a very successful tool – by the end of the campaign more than 1.5 million Quick Donate users donated $115 million – and so the team was anxious to find ways to get more donors to sign up for Quick Donate.

Figure 3 shows two versions of the page that donors saw just after making their donation. The control page asked customers: "save your payment information for next time," while the treatment page made it seem as if saving the payment information was part of the current process by saying: "now, save your payment information." When users were randomly assigned to the two treatments, the percentage of customers who saved their payment information was 21% greater among those who saw the *segue copy*.

This example raises a key issue that testing managers face in practice: how to measure the effect of the treatment. In this case, the team chose to compare treatments based on how many customers signed up for the Quick Donate program, and this is a logical choice as that is the immediate goal of these marketing treatments. However, Quick Donate sign-ups do not result in an immediate monetary gain for the campaign. One might also legitimately prefer to compare these two treatments based on how many actual donations are received in the subsequent month for those in each group, although this would require more time and tracking capability to measure effectively.

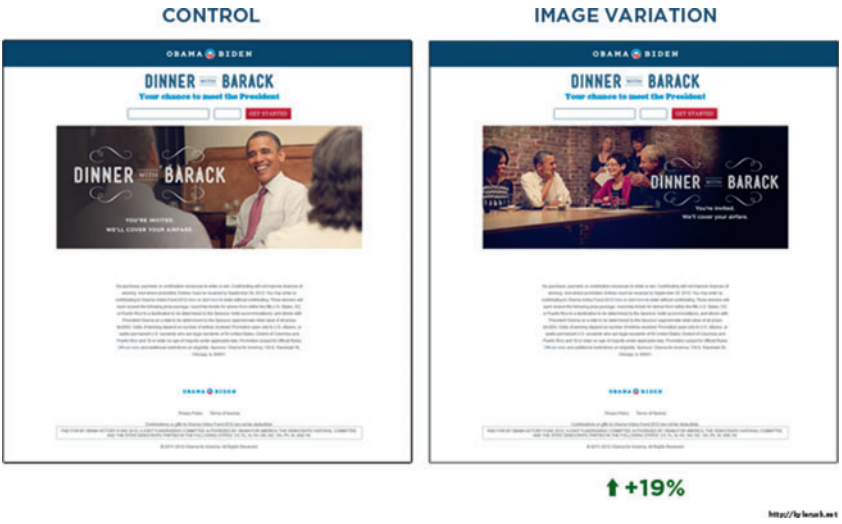

Fig. 2 Image test for Obama campaign (Adapted from Rush 2012b)

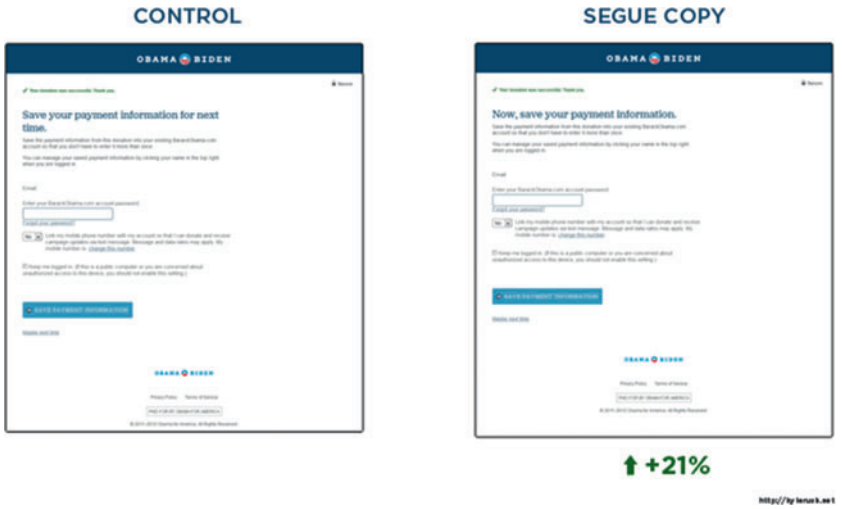

Fig. 3 Copy test for Obama campaign (Adapted from Rush 2012b)

In field experiments in digital marketing, it is common to measure a variety of outcomes within the same experiment, both those that are directly related to the short- and long-term effects of the treatment and potential side effects such as increased costs or increased complaints. (Medical experiments face a similar challenge in defining response measures: in testing a new cancer treatment, researchers must decide whether to compare treatments based on a near-term outcome such as

the recurrence of cancer in the subsequent 5 years or a longer-term outcome such as mortality in the next 20 years.)

The straightforward randomization and measurement available on the web platform allow for easy causal interpretation of the results, which in turn makes it easy for decision makers to act immediately on the findings without much risk of paralysis by analysis. As Rush describes, "In looking at the overall results I think you could say our efforts paid off. We increased donation conversions by 49%, sign up conversions by 161% and we were able to apply our findings to other areas and products." And this sort of result is not unique: spurred on by a number of popular business books with titles like *Always Be Testing* (Eisenberg and Quarto-von Tivadar 2009), *Experiment!* (McFarland 2012), and *A/B Testing* (Siroker and Koomen 2013) where other examples of the Obama campaign's optimization are reported, many firms are finding ways to making field experiments a regular part of how they make decisions.

Field Experiments in the Academic Literature

Field experiments are becoming popular as a tool for exploring marketing theory (Gneezy 2017), and there are many online and offline field experiments reported in the academic literature.

Offline field experiments can, for instance, be run in retail stores like Chen et al. (2012) did to test how different types of promotions can impact the volume of purchases. They tested whether the bonus pack or an equivalent price decrease of a product has an impact on the sales figures changing the promotion type on a weekly basis for 16 weeks. The employment of only one store allowed keeping all external factors constant (e.g., store layout, employees, background of customers, neighboring environment), increasing internal validity at the expense of external validity.

Furthermore, field experiments are often conducted over a long period of time in order to identify long-term effects. For example, Bawa and Schoemaker (2004) conducted two field experiments each one over a 2-year time frame aimed at estimating the long-run effect of free sampling on sales. In both cases, they recorded the sales data of the customers over 1 year (panel data). After delivering the sample at the end of the first year, the volumes were registered for another year. Of course, the longer the time frame, the higher the probability that external factors can influence the participants. In general, problematic marketing-related extraneous factors depend on the respective context and on the research topic.

Online-controlled experiments have gained popularity because of the increased digitalization of companies that are more and more engaging in a test and learn mentality. As we have discussed, A/B tests can easily be implemented to examine how users react to different webpage layouts and designs. An example is Yang and Ghose (2010), who measured the impact of different search advertising strategies on the click through rate, conversion rate, and revenues. All of these measures give an indication of how the customers use the website.

A study revealing how the use of field experiments can shed new light on existing and well-established theories is the recent paper by Anderson and Simester (2013). Standard models of competition predict that firms will sell less when competitors

target their customers with advertising. This is particularly true in mature markets with many competitors that sell relatively undifferentiated products. However, the authors present findings from a large-scale randomized field experiment that contrast sharply with this prediction. The field experiment examines the effect of competitors' advertising on sales at a private label apparel retailer. To examine this effect, the researchers sent competitive advertisement mailings to the treatment group. As customers normally have no comparison of whether other people receive the same or different mailings, they do not realize that they are part of an experiment. Results show that, surprisingly, for a substantial segment of customers, the competitors' advertisements increased sales at this retailer.

Recommended readings for those interested in online advertising are the field tests employed by Goldfarb and Tucker (2011a, c). In the same area, Blake et al. (2015) and Kalyanam et al. (2015) published large-scale field experiments aimed at studying the causal effectiveness of paid search ads. They find somewhat contradictory results: Blake et al. (2015) showed that returns from paid search ads for eBay are minimal, while Kalyanam et al. (2015) find that search ads are effective for other retailers. In a recent working paper, Simonov et al. (2015) have also confirmed that search advertising does have some benefit for less-well-established brands. They use a large-scale, fully randomized experiment on Bing data studying 2500 brands. These experiments rely on treatment and control groups made up of various geographic regions where advertising can be turned on or off; using such *geo-experiments* to measure ad effectiveness has also been suggested by researchers at Google (Vaver and Koehler 2011).

Randomized holdouts take this idea of non-exposure to customer-level experiments and are rapidly becoming popular in many industries. In a randomized holdout experiment, the marketer selects a group of customers at random to not receive planned marketing communication, such as an email, a catalog, or a promotional offer. Comparing the treated and the holdout group allows the marketer to make a causal measurement of the treatment effect, i.e., the incremental sales lift of the marketing. Hoban and Bucklin (2015) report on randomized holdout experiments in display advertising, Zantedeschi et al. (2016) report on randomized holdouts for catalog and email campaigns, and Sahni et al. (2015) report on randomized holdouts for discount offers. All of these studies find positive incremental effects of marketing. However, Lewis and Rao (2015) report similar experiments on display advertising and find effect sizes that are so small that it would be difficult to accurately measure the returns on advertising.

Lambrecht and Tucker (2013) run a field experiment with an online travel firm to examine whether *dynamic retargeting*, a new form of personalized advertising that shows consumers ads that contain images of products they have looked at before on the firm's own website, is more effective than simply showing generic brand ads. Even if this new strategy integrates the usage of both internal and external browsing data, results revealed that dynamic retargeted ads are on average less effective than traditional retargeting.

Ascarza et al. (2016) analyze retention campaigns based on pricing plan recommendations, and the results emerging from their field experiment surprisingly show

that being proactive and encouraging customers to switch to cost-minimizing plans can increase rather than decrease customer churn.

As the MSI-Tier 1 priorities suggest, the customer journey is developing into a multimedia, multiscreen, and multichannel era (mobile = physical + digital worlds). Considering multichannel customer management literature, Montaguti et al. (2016) test the causal relationship between multichannel purchasing and customer profitability. Within a field experiment, they show that multichannel customers are indeed more profitable than they would be if they were single-channel customers providing insights on how multichannel shopping leads to higher profit.

Andrews et al. (2015) had the opportunity to collaborate with one of the world's largest telecom providers managing to gauge physical crowdedness in real time in terms of the number of active mobile users in subway trains. Their research examines the effects of hyper-contextual targeting with physical crowdedness on consumer responses to mobile ads, and results based on a massive field experiment counting a sample of 14,972 mobile phone users suggest that, counterintuitively, commuters in crowded subway trains are about twice as likely to respond to a mobile offer by making a purchase vis-à-vis those in non-crowded trains.

Dubé et al. (2015) implemented another massive field experiment to test an information theory of prosocial behavior. A long literature in behavioral economics has generated a collection of empirical examples where economic incentives counterintuitively reduce the supply of prosocial behavior. The data comes from two field experiments involving a consumer good bundled with a charitable donation. Considering a population of 15 million subscribers living 2 km from a theater and who purchased a ticket via phone in the previous 6 months, the sample consisted of 4200 randomly chosen individuals. Results suggest that price discounts crowd out consumer self-inference of altruism.

Nevertheless, the aforementioned papers are only some of those interesting works published involving the use of field experiments. We leave to the reader's curiosity the task to look for other field experiments!

Conclusions

As can be seen from the previous section, there are numerous examples of both companies and academics using field experiments to answer tactical questions and test marketing theory. The increasing use of field experiments in marketing is also enhancing the collaboration between firms and academia. The big challenge and opportunity here are the reconciliation of academics doing "big stats on small data" with practitioners doing "small stats on big data."

This chapter has laid out the key ideas one should think about when designing field experiments. For the reader interested in more detail, a major author of reference is John A. List, who focuses on field experiments in economics. In List (2004) the author presents a series of field experiments he conducted about theories of discrimination, and in a slightly more recent paper (2006), he reviews a broad set of field experiments to explore the implications of behavioral and neoclassical

theories as well as of topics ranging from the economics of charity to the measurement of preferences. Furthermore, in 2011 he proposed 14 tips to follow for improving academic's chances of executing successful field experiments in collaboration with companies. We suggest practitioners to refer to this checklist, before implementing their experiment ideas.

Of course, it is unavoidable to meet some challenges in the implementation and use of field experiments. First of all, as pointed out by Levitt and List (2009), field experiments do not provide the same extent of control as laboratory experiments. Therefore, internal validity is often lower, and, because of a lower level of control, potential confounding variables should be identified before starting and recorded during the experiment in order to control for them using statistical methods (Homburg 2015; Gneezy 2017). Pre-testing and continuous monitoring during the experiment are helpful to identify excluded effects and record general trends like a change of the general market conditions which can impact the sales volume independently of the experiment (Gerber and Green 2012; Gneezy 2017). This issue further reveals that researchers should put much effort and time into the planning stage and in the experimental design. On top of that, a relatively high level of knowledge of the whole experimental design and of the underlying constructs is required upfront (Levitt and List 2009). Other challenges concern privacy and security regulations that unavoidably tend to limit collection/retention of data (Goldfarb and Tucker 2011b). Future researchers should focus on the development of analytics that can overcome such limitations and on the proactive development of methods for protection of customer privacy.

In summary, this chapter outlines and argues that field experiments are, next to big data analytics, one of the major advances of the digital age which allow firms to reveal the causality between two processes, actions or observations. Managers and researchers have now to accept the challenge by ensuring that the causal inferences of their field experiments are both correct and useful in terms of advancing management and marketing practice. We hope this chapter encourages and helps managers in considering field experiments as a state-of-the-art market research approach for collection, analysis, and interpretation of market-related information.

Cross-References

- ▶ Analysis of Variance
- ▶ Experiments in Market Research

References

Aaker, D. A., Kumar, V., Day, G. S., & Leone, R. P. (2011). *Marketing research*. Hoboken: Wiley.

Anderson, E. T., & Simester, D. (2013). Advertising in a competitive market: The role of product standards, customer learning, and switching costs. *Journal of Marketing Research, 50*(4), 489–504.

Andrews, M., Luo, X., Fang, Z., & Ghose, A. (2015). Mobile Ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science, 35*(2), 1–17.

Ascarza, E., Iyengar, R., & Schleicher, M. (2016). The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. *Journal of Marketing Research, 53*(1), 46–60.

Bawa, K., & Shoemaker, R. (2004). The effects of free sample promotions on incremental brand sales. *Marketing Science, 23*(3), 345–363.

Blake, T., Nesko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *Econometrica, 83*(1), 155–174.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–312.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Chen, H., Marmorstein, H., Tsiros, M., & Rao, A. R. (2012). When more is less: The impact of base value neglect on consumer preferences for bonus packs over price discounts. *Journal of Marketing, 76*(4), 64–77.

Crook, T., Brian, F., Ron, K., & Roger, L. (2009). *Seven pitfalls to avoid when running controlled experiments on the web*. In Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining.

Dubé, J.-P., Luo, X., & Fang, Z. (2015). *Self-signaling and pro-social behavior: A cause marketing experiment*. Fox school of business research paper no. 15-079. Available at SSRN: http://ssrn.com/abstract=2635808 or http://dx.doi.org/10.2139/ssrn.2635808

Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach*. Cambridge: Cambridge University Press.

Eisenberg, B., & Quarto-von Tivadar, J. (2009). *Always be testing: The complete guide to Google website optimizer*. New York: Wiley.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver and Boyd.

Gerber, A. S., & Green, D. P. (2008). Field experiments and natural experiments. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology*, *Oxford handbooks online* (pp. 357–381). Oxford: Oxford University Press.

Gerber, A. S., & Green, D. P. (2012). *Field experiments. Design, analysis, and interpretation*. New York: Norton.

Gneezy, A. (2017). Field experimentation in marketing research. *Journal of Marketing Research, 46*, 140–143.

Goos, P., & Jones, B. (2011). *Optimal design of experiments: A case study approach*. New York: Wiley.

Goldfarb, A., & Tucker, C. E. (2011a). Online display advertising: Targeting and obtrusiveness. *Marketing Science, 30*(3), 389–404.

Goldfarb, A., & Tucker, C. E. (2011b). Privacy regulation and online advertising. *Management Science, 57*(1), 57–71.

Goldfarb, A., & Tucker, C. E. (2011c). Advertising bans and the substitutability of online and offline advertising. *Journal of Marketing Research, 48*(2), 207–227.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009–1055.

Hoban, P. R., & Bucklin, R. E. (2015). Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research, 52*(3), 375–393.

Homburg, C. (2015). *Marketingmanagement. Strategie – Instrumente – Umsetzung – Unternehmensführung. Lehrbuch*. Wiesbaden: Springer Gabler.

Homburg, C., Kuester, S., & Krohmer, H. (2013). *Marketing management. A contemporary perspective*. London: McGraw-Hill Higher Education.

Iacobucci, D., & Churchill, G. A. (2010). *Marketing research. Methodological foundations*. Mason: South-Western Cengage Learning.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social and biomedical sciences: An introduction*. New York: Cambridge University Press.

Kalyanam, K., McAteer, J., Marek, J., Hodges, J., & Lin, L. (2015). *Cross channel effects of search engine advertising on brick and mortar retail sales: Meta analysis of large scale field experiments on Google.com*. Working paper.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery, 18*(1), 140–181.

Koschate-Fischer, N., & Schandelmeier, S. (2014). A guideline for designing experimental studies in marketing research and a critical discussion of selected problem areas. *Journal of Business Economics, 84*, 793–826.

Landsberger, H. A. (1958). *Hawthorne revisited*. Ithaca: Cornell University.

Lambrecht, A., & Tucker, C. E. (2013). When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research, 50*(5), 561–576.

Ledolter, J., & Swersey, A. J. (2007). *Testing 1-2-3. Experimental design with applications in marketing and service operations*. Standford: Standford University Press.

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review, 53*(1), 1–18.

Lewis, R. A., & Rao, J. M. (2015). The unfavourable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 1941–1973.

List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *Quarterly Journal of Economics, 119*(1), 48–89.

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives, 25*(3), 3–16.

McFarland, C. (2012). *Experiment! Website conversion rate optimization with A/B and multivariate*. Berkeley: New Riders.

Montaguti, E., Neslin, S. A., & Valentini, S. (2016). Can marketing campaigns induce multichannel buying and more profitable customers? A field experiment. *Marketing Science, 35*(2), 201–217.

Neyman, Jerzy. (1923[1990]). On the application of probability theory to agricultural experiments: Essay on principles. Section 9. *Statistical Science*, 5 (4), 465–472. Translated by Dabrowska, Dorota M. and Terence P. Speed.

Rush, K. (2012a). *Meet the Obama campaign's $250 million fundraising platform*. Blog post 27 Nov 2012.

Rush, K. (2012b). *Optimization at the Obama campaign: a/b testing*. Blog post 12 Dec 2012.

Sahni, N., Dan, Z., & Pradeep, C. (2015). *Do targeted discount offers serve as advertising? Evidence from 70 field experiments*. Stanford University Graduate School of Business research paper no. 15-4. Available at SSRN: http://ssrn.com/abstract=2530290 or http://dx.doi.org/10.2139/ssrn.2530290

Sándor, Z., & Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research, 38*(4), 430–444.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont: Wadsworth Cengage Learning.

Simonov, A., Nosko, C., & Rao J. M. (2015). *Competition and crowd-out for brand keywords in sponsored search*. Available at SSRN: http://ssrn.com/abstract=2668265 or http://dx.doi.org/10.2139/ssrn.2668265

Siroker, D., & Pete K. (2013). A/B testing. Wiley.

Teele, D. L. (2014). *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. New Haven & London: Yale University Press.

Vaver, J., & Koehler, J. (2011). *Measuring ad effectiveness using geo experiments*. Google Research working paper.

Yang, S., & Ghose, A. (2010). Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science, 29*(4), 602–623.

Zantedeschi, D., McDonnell Feit, E., & Bradlow, E. T. (2016). Modeling multi-channel advertising response with consumer-level data. *Management Science*, Articles in Advance. http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2451