Field Experiments in Marketing

Anja Lambrecht and Catherine E. Tucker*

September 10, 2015

Abstract

In a digitally enabled world, experimentation is easier. Here, we explore what this means for marketing researchers, and the subtleties of designing field experiments for research. It gives guidelines for interpretation and describe the potential advantages and disadvantages of this methodology for classic areas of marketing.

Keywords: Field Experiments, Field Tests, Causal Inference

*Anja Lambrecht is an Assistant Professor of Marketing at London Business School, United Kingdom. Catherine Tucker is a Professor of Management Science at the MIT Sloan School of Management, Cambridge, MA. and Research Associate at the NBER. All errors are our own.

The digital revolution has led to an explosion of data for marketing. This 'Big Data' available to researchers and practitioners had created much excitement about potential new avenues of research. In this chapter, we argue that an additional large and potentially important part of this revolution is the increased ability for researchers to use data from field experiments facilitated by digital tools.

Marketing as a field, perhaps because of its historical relationship with psychology, has embraced and idealized field experiments from an early stage in its evolution. Roberts (1957), when evaluating statistical inference as a tool for Marketing Research, wrote the following still powerful passage on the merits of field experiments:

> In experimental applications, managerial actions are actually tried out with the aim of discovering the responses to these actions. All other applications are nonexperimental or 'observational.' [...]
>
> The key to modern statistical design of experiments is withholding experimental stimuli at random. To the extent that randomization and the other conditions above are met, the responses actually observed will reflect the 'true' effects of the stimuli plus random or chance variation. Statistical procedures then need cope only with the interpretation of chance variation.

In other words, marketing research has from the beginning drawn a clear and favorable line between experimental techniques which allow a causal interpretation and everything else. Therefore, we emphasize that the aim of this chapter is not to claim any novelty in our procedural guide to the use of field experiments in marketing research, but instead to attempt to update these techniques for a digital world that has made their implementation easier, and to provide a guide to the pitfalls of such techniques for researchers who are new to them.

In this chapter, we set out the field experiment methodology and its main advantages

and also lay out some general guidance for the interpretation of statistical results from field experiments. We then consider various applications of field experiments to marketing. We then conclude by emphasizing the limitations to this methodology.

1 A Description of Field Experiment Methodology

In this section, we describe why field experiments are useful from a statistical point of few and five steps that researchers need to reflect upon when designing a field experiment and interpreting its results. The focus of this chapter is field experiments or interventions in the real world, rather than the laboratory. The Lee and Tybout chapter in this volume discusses the lap experiment method and we encourage interested readers to read that chapter for more information.

1.1 Why a field experiment?

The *raison d'etre* of a field experiment is to provide causal inference. List (2011) (p. 8), in his justification of the use of field experiments, puts it well when he says that "The empirical gold standard in the social sciences is to estimate a causal effect of some action." Therefore, it is useful for marketing researchers to understand the econometric framework, upon which basis field experiments make their claim to provide causal inference that is superior to other techniques.

A useful approach is that of 'potential outcomes' (Rubin, 2005).[1] In this approach, for any treatment (x), each i has two possible outcomes:

- y_{i1} if the individual i experiences x
- y_{i0} if the individual i does not experience x

The difference between y_{i1} and y_{i0} is the causal effect. However, this is problematic to measure, because a single individual i cannot both receive and not receive the treatment.

[1]This builds on a large number of books and articles that have covered similar material (Angrist and Pischke, 2009; Manski, 2007; Meyer, 1995; Cook and Campbell, 1979; Imbens and Wooldridge, 2009)

Therefore, only one outcome is observed for each individual. The unobserved outcome for any individual is the "counterfactual." The lack of observable counterfactuals for each individual means that those who experience x and those who do not are different, even if there is a field experiment. Instead, a field experiment ensures that *ex ante*, via random assignment, any differences between the treated and control group should not matter.

1.2 Step 1: Decide on unit of randomization

The above framework makes the motivation for the use of field experiments straightforward. However, the term 'random assignment' and its implementation turn out to be far more challenging than they appear in this theoretical setting. Before random assignment can occur, the researcher needs to decide at what degree of granularity random assignment should occur. Theoretically, randomization could happen, for example, at the level of the individual, household, town, website, store or firm. Often, this choice of the 'unit of randomization' will determine the success of a field experiment in terms of statistical power as well as how convincing the results of the field experiment are.

At the highest level of generality, the statistical power of a randomized experiment is likely to increase with greater granularity of the unit of randomization. To consider why, contemplate the following scenario: Imagine a firm selling bottled water wants to use a field experiment to test different pricing strategies. It decides (at random) to test 'every day low pricing' west of the Mississippi and 'hi-lo' pricing east of the Mississippi. In other words, there are just two units - in this case geographical clusters of stores - that are randomized. Imagine too, that a drought hits the territory west of the Mississippi at the same time as the experiment. Then, even if every day low pricing appears to be selling more bottled water, it is not clear whether this was due to the randomized experiment or to the drought. Put differently, the lack of granularity in randomization reduced the chance that *ex ante* the 'unobserved ways' do not matter, as this lack of granularity also made it more likely that

there might be systematic error associated with one territory.[2]

Given this challenge, a researcher might always think that it would be attractive to choose the most granular unit of randomization technologically possible. However, there are also two constraints that argue against granularity. First, there are the constraints imposed by the costs and logistics of having a finely-grained unit of observation. Second, the researcher needs to minimize the potential for spillovers and crossovers between experimental treatments.

In a non-digital environment, randomization is often constrained simply by the ability to identify an individual and deploy a randomization algorithm. However, the digital environment makes the conduct of very granular field experiments straightforward and easy. The ease of such a procedure has led to a new managerial language of 'split tests' or 'a/b testing'; commercial firms such as Optimizely[3] now allow managers to independently and easily run field tests to evaluate the effects of different landing pages or website content using the highly granular unit for randomization of an individual website visit.

However, in an offline environment maintaining more units for randomization could potentially still be very costly or logistically difficult. For example, suppose a researcher wanted to evaluate the effect of different 'sales scripts' on the performance of a sales department. Potentially, it might be attractive to randomize which sales script was used for each call. However, practically and logistically it might be simpler and cheaper if instead each sales person would be randomly allotted to perform a single sales script when making calls. This would reduce training costs and organizational complexity. However, it introduces the risk of systematic bias if, for example, more able sales people were accidentally randomized into one condition rather than another. Of course it is possible to use stratified randomization if such ability is observable in advance, but potentially it may not be.

[2]One way of dealing with this possibility when there is data on the observable characteristics of different units is stratified randomization which we discuss next.

[3]`optimizely.com`

1.3 Step 2: Ensure no spillover and crossover effects

A more pressing problem, however, than one of simple costs or logistical complexity when it comes to choosing the right unit of randomization, is the need to minimize spillovers and crossovers between experimental treatments. A spillover occurs when a treated individual (or other unit) affects the outcomes for other untreated individuals.[4] Suppose a firm randomly selected an individual to receive a free mobile phone. Potentially their adoption of a mobile phone could affect the adoption outcomes of their relatives and friends, even if those relatives and friends were supposedly untreated. If such spillovers are a large concern, then one way of addressing them would be to randomize at the level of plausibly isolated social networks such as a community, rather than randomizing at the level of the individual.[5]

A crossover occurs when an individual who was supposed to be assigned to one treatment is accidentally exposed to another treatment. Suppose, for example, a canned soup company is testing different advertising messages in different cable markets, and an individual is exposed to a different advertising message from that of their home market because they are travelling. This could potentially lead to mismeasurement of the treatment, especially if there were systematic patterns in travel which led such crossovers to not simply be random noise. Indeed, this is one issue we faced even in a digital context in Lambrecht and Tucker (2013), where randomization was implemented on an individual-day level rather than at the level of the individual. When an individual arrived at a website, a digital coin-toss determined whether they were exposed to a personalized ad, taking no account of what type of ad the individual had previously been exposed to. So an individual could be placed into different conditions on different days, and the number of different conditions they were placed into was itself related to their frequency of website use. Here, we took care to include appropriate

[4]Roberts (1957) puts this well by advising the researcher to make sure that *The population being studied can be broken down into smaller units (families, stores, sales territories, etc.) for which the experimental stimuli can be measured and for which responses to the stimuli are not 'contagious.'*

[5]Such spillovers are currently attracting the attention of econometricians at the frontier of the analysis of randomized experiments. We point the interested reader to the work of Barrios et al. (2012), among others.

control variables, but this potential for crossover between advertising conditions could have been addressed in the experimental design if the firm we were working with had randomized at a less granular level.

1.4 Step 3: Decide on complete or stratified randomization

The second question that a researcher should tackle after establishing the unit of randomization is whether to conduct stratified randomization or complete randomization. In complete randomization, individuals (or the relevant unit of randomization) are simply allocated at random into a treatment. In stratified randomization, individuals are first divided into subsamples based on covariate values so that each of the subsamples are more homogenous relative to that covariate than the full sample. Then, each individual in each of these subsets is randomized to a treatment.[6] This stratified technique is useful if a covariate is strongly correlated with an outcome. For example, household income may be strongly correlated with purchase behavior towards private label brands. Therefore, it may make sense, if the researcher has access to household-level data, to stratify the sample prior to randomization to ensure sufficient randomization occurs within, for example, the high-income category.

There is a relatively large empirical literature discussing the merits of different approaches to stratification in the context of schooling experiments and experiments within the developing world. For examples of this debate, see Bruhn and McKenzie (2008) and Imai et al. (2008, 2009). It is worth pointing out, though, that the typical school setting on which this debate is focused is often less relevant to marketing applications. First, often in marketing it is hard to collect reliable data before an experiment which would allow stratification and subsequent random assignment before the experiment. Second, much of the debate is motivated by experimental treatments such as a change in school class size which are very costly and therefore obtaining statistical efficiency from a small number of observations is paramount.

[6] A special case of a such a stratified design is a pairwise design where each stratum contains a matched pair of individuals, one of whom receives the treatment and the other does not.

For example, when randomizing 30 different schools into different class size conditions, one might not obtain any statistical precision in estimates simply because by unlucky chance the richest schools were all randomly allocated into the lowest class size condition. However, for many marketing applications such as pricing or advertising, the kind of cost constraints which would restrict the researcher to only look at only 30 units of observations are less likely to be present. Furthermore, reliable data which would allow such stratification may not be present.

1.5 Step 4: Ensure that appropriate data is collected

After ensuring that randomization is appropriate, researchers should carefully consider what type of data they need for their later analysis and ensure the practical set-up allows them to collect this data. This is particularly important in digital environments where different parties have access to different types of data and it is not always obvious how these can be collected and linked. For example, advertising networks have access to ad exposure data but it may require additional steps to ensure that they likewise capture purchase data and can link those to ad exposures. In Lambrecht et al. (2015), we were unable to provide this link. By contrast, in Lambrecht and Tucker (2012) we worked with the web hosting provider conducting the field experiment to implement Google Analytics to track consumers arriving from Google's search engine at the website of the web hosting provider. Additionally, researchers should carefully consider data points that are not directly linked to measuring the outcome of the randomization, but thay may help the researcher understand the behavioral mechanism or rule out alternative interpretations. For example, while conducting a field experiment on Twitter, Lambrecht et al. (2015) concurrently collected data from an independent source, on the size of all Twitter trends their study was focusing on, on every day of the field experiment from an additional, independent source. This data served to later rule out that the size of the trends studied led to the effect of interest.

Any researcher interested in field experiment techniques should be aware of the potential need for a large sample when conducting a field experiment, especially when the magnitude and direction and heterogeneity of the treatment effect is unknown.[7] It is devastating to run a field experiment and obtain statistically imprecise estimates of the causal effect due to lack of sample size. There are many settings where this may be a concern. For example, Lewis and Rao (2013) show that for many online advertising campaigns the effect is so small and heterogeneous that measurement even with millions of observations can result in imprecise estimates. It may be possible to identify such contexts by reference to the explanatory power of different variables in prior observational (and non-randomized studies). In general, though it is difficult to give practical advice to researchers beyond aiming for as expansive a sample and data collection effort as possible.

1.6 Step 5: Interpreting results from a field experiment

Though in theory, the 'potential outcomes' approach means that interpretation should be straightforward, in practice there are numerous issues that the researcher should be aware of when interpreting their statistical results. In general, the key issue is understanding exactly what is different between the groups who were treated and those who were not, and being careful about how to generalize this difference.

A key consideration for researchers is how the length of time the field experiment ran for will affect their interpretation of their results.[8] Anderson and Simester (2004) highlighted the importance of making sure the researcher has access to a long enough period of data by showing that the long-run effects of promotional depth were negative for established customers, though in the short run they could look deceptively attractive due to their ability to attract new customers. In general, researchers should try and collect data for as long

[7] Roberts (1957) states that *The sample size is large enough to measure important responses to experimental stimuli against the background of uncontrolled sources of variation.*

[8] Roberts (1957) urges researchers to ensure that *The experiment is run sufficiently long that responses to experimental stimuli will have time to manifest themselves.*

a period as possible to understand whether any treatment they measure is stable, dissipates or increases in its effect over time. However, for many field experiments it is hard to measure long-run effects as the researcher does not have the ability to monitor treated and untreated individuals over time. Therefore, in most settings researchers should carefully consider whether the causal effect they establish truly reflects the long-run treatment effect.

The existence or importance of Hawthorne effects, where the mere fact of being observed as part of a field experiment can alter outcomes, is the subject of much academic debate (Parsons, 1974; Adair, 1984; Jones, 1992; McCarney et al., 2007).[9] In general, however, this kind of critique invites a researcher to be thoughtful about what really is the difference between the 'treatment' and the 'control' and what specifically they measure. The researcher should provide reassuring evidence for the reader that the causal effect they measure between the treatment and control is associated with the part of the treatment they claim it is. For example, Burtch et al. (2015) use data from a field experiment which introduced new privacy settings in a crowdfunding setting. They devote much space in their article to giving the reader evidence that the change they measure in crowdfunding propensity really was a result of the change in privacy setting rather than simply the introduction of a new screen or potential navigation costs for the website user.

One obvious concern that researchers face, especially those who work with firms, is that there may be compromises or challenges to randomization. Firms may only be willing, for example, to experiment with, in their view, less successful media or sales territories, and unwilling to experiment with more successful ones. Similarly, firms may only be willing to incur the costs of experimentation for their best customers. Simester et al. (2009) provides a nice example of how a researcher faced with such constraints can describe the selection criteria which constrained randomization and provide reassuring evidence and discussion to

[9]Roberts (1957) emphasizes that researchers should try and make sure *'Neither the stimulus nor the response is changed by the fact that an experiment is being conducted.'*

allow the reader to understand what the constraints mean. In their particular case, they used the company's decision to distinguish between 'best' customers and 'other' customers when determining random assignment as a useful way of exploring the underlying behavioral mechanism. In general, though, in such circumstances the key procedure for any researcher is to be upfront about the limitation and discuss its implications for generalizability.[10]

2 What Marketing Contexts Can Use Field Experiments?

2.1 Promotion and Marketing Communications

Marketing communications, and especially advertising, is perhaps the area that has been most revolutionized by the ability to conduct field experiments in the digital space.

Some work has focused on measuring the effectiveness of different forms of advertising. Lewis and Reiley (2014b) measure the effect of online advertising on offline sales and find positive effects. Draganska et al. (2014) use field test data to compare the effectiveness of television and internet advertising. Blake et al. (2014) examine the impact of paid search advertising on purchases in a large-scale field experiments at eBay. Sahni (2011) studies how the different timing of ads moderates their effectiveness. Offline, Bertrand et al. (2010) measure the effectiveness of advertising in the developing world.

Other work has used field experiments to measure the effectiveness of advertising for different kind of users and product contexts, such as older internet users (Lewis and Reiley, 2014a) and different kinds of products (Bart et al., 2014). Yet another way in which field experiments can be useful in the context of marketing communications is to explore which groups of consumers are most responsive to targeted ads. Lambrecht et al. (2015) show that early trend propagators are on average less responsive to promoted tweets, that is advertising messages on Twitter, than consumers who post on the same trends later on. Hoban and

[10]Roberts (1957) somewhat anticipates this when he urges researchers to ensure that *'The experimentor is able to apply or withhold, as he chooses, experimental stimuli from any particular unit of the population he is studying.'*

Bucklin (2014) find that users in most stages of the purchase funnel are receptive to ads but not those who previously visited the site without creating an account.

Researchers have also used digital experiments to explore optimal ad content and design. Fong (2012) explores the content of targeted email offers and find that a closely matched offer may weaken a customer's incentives to search beyond the targeted items. Lambrecht and Tucker (2012) explore how consumers respond to different prices advertised in Google search ads. Ascarza et al. (2015) find that customers who were randomly offered recommendations as to their mobile phone plan were more likely to churn than those who were not offered recommendations.

Much of this literature has emphasized that not all digital enhancements of ad content are positive. Aral and Walker (2011) show that viral ad design is only of limited success. Goldfarb and Tucker (2011a) show that there is a tradeoff between the level of targeting of a display ad's content and the ad's intrusiveness. Goldfarb and Tucker (2014) show that there is a tradeoff between the degree of standardization of digital ad formats and how effective they are at attracting viewers' attention – for most ads, recall of banner advertising declines the more ads conform to standard formats, especially for ads that focus on brand logos, and less so for ads designed by advertising agencies. Tucker (2014a) shows that social endorsements are only of limited effectiveness in enhancing ad content. Lambrecht and Tucker (2013) show that very personalized ad product content can backfire unless a consumer's browsing history indicates that they have reached a stage in their purchase process where they are ready to buy.

One of the challenges of optimizing online advertising is identifying and implementing optimal policies in real time. Schwartz et al. (2013) solve the problem of maximizing customer acquisition rates by testing many ads on many websites while learning which ad works best on each website by implementing a multi-armed bandit policy that adjusts in real time in a large adaptive field experiment.

2.2 Pricing

Firms and researchers can use field experiments to understand consumer response to different prices and set optimal prices. Offline, Anderson and Simester (2003) looked at the effect of $9 price endings and Anderson and Simester (2001) show that sale signs are less effective the more products have them.

The effect of promotions on sales has attracted much attention in both offline and online settings. Anderson and Simester (2010) extend earlier work to show that discounts can lead to customer antagonism, especially among loyal customers. Lee and Ariely (2006) report on a series of field experiments in a convenience store where consumers were randomly exposed to different treatments such as varying when during the shopping process conditional coupons (of the form 'Spend $X and get $1 off') were handed to them and the amount of the coupon. They find that conditional coupons are more effective in influencing consumers spending when consumers goals are less concrete. Sahni et al. (2014) find a positive effect of promotions that largely comes not from redemption of the offers but from a carryover to the following week. Their study also highlights, however, that higher risks of crossover and spillover effects exist when experimenting with prices online, especially when price differences between test conditions become large and social networks are prevalent. Fong et al. (2015) and Andrews et al. (2015) are among a recent body of work exploring the effectiveness of mobile promotions.

While a majority of field experiments focus on B-to-C settings, a study by Tadelis and Zettelmeyer (2011) demonstrates that field experiments can likewise be very useful in understanding in B-to-B transactions. The authors examine in a large-scale field experiment that randomly discloses quality information in wholesale automobile auctions how information disclosure affects auction outcomes.

Last, field experiments have served to understand consumers' response to pay-what-you-

want pricing. Kim et al. (2009) find in multiple field studies that prices paid are significantly greater than zero and can even increase revenues. These studies rely on experimentation over time, highlighting the difficulty for offline stores, here restaurants, to concurrently implement different pricing mechanisms. By contrast, Gneezy et al. (2012) randomized in several field experiments the price level and structure to which consumers were exposed. They show that often, when granted the opportunity to name the price of a product, fewer consumers choose to buy it than when the price is fixed and low. Jung et al. (2014) demonstrate that when asked to pay as much as they like, merely reframing payments to be on behalf of others, not their own, leads people to pay more. Broadly related, Gneezy et al. (2010) show that a charitable component in a purchase increased sales significantly when coupled with a 'pay-what-you-want' pricing mechanism.

2.3 Product

It can be challenging to implement field experiments to better understand the relative performance of alternative new products, designing new products or testing them relative to the competition. In many industries, operational constraints prevent firms from launching different product alternatives concurrently, especially in the non-digital economy where such field experiments can be very costly. In addition, experimenting with products can confuse customers and lead to spillover and crossover effects. It may also lead to competitive response prior to a full-scale product introduction.

One potential avenue for researchers is to work with firms who already test the profitability of new products and their effect on existing product lines. For example, MacDonalds regularly tests new menu items by rolling out a new product to a small subset of stores.[11] Additionally, there are possibilities for field experiments regarding products in the developing world. For example, using the example of antimalarial bed nets, Dupas (2014) show that

[11] http://www.mcdonalds.co.uk/ukhome/whatmakesmcdonalds/questions/food/nutritional-information/how-do-you-product-test-new-products.html

rather than deterring future purchases, one-off subsidies can actually encourage willingness to pay.

Additionally, researchers have used field experiments to better understand customer needs in the design of new products, product customization, and presentation of product information. Boudreau et al. (2011) shows the possibility of using field experiment techniques in product design using data on software contests. Hildebrand et al. (2014) find that customers who were randomly designed to a condition where they would create a customized product from a starting solution are more satisfied with their purchase than customers who are assigned to a condition that requires an attribute-by-attribute configuration. Relatedly, Levav et al. (2010) demonstrate in a field experiment that when consumers customize products, the order in which attributes are presented changes their revealed preferences. When users of a social networking site can choose product characteristics, Sun et al. (2012) find that subjects were more likely to diverge from the popular choice among their friends as the popularity of that choice increased.

A broadly related question is how consumers respond to different information provided in search results. Nosko and Tadelis (2015) implement a field experiment where they change the search results for a randomly chosen subset of buyers on eBay using a new suggested measure of quality. They find that their suggested measure of quality increases the quality of transactions and, consequently, the retention of buyers.

2.4 Distribution

Distribution decisions often involve conflicts of interest, are long-term, are difficult to change and costly to implement. As a result the use of field experiments tends to be difficult. However, digital technology and specifically the online channel open up new avenues for researchers.

Though there are few field experiments focused on channels, we highlight a subset papers

that use natural experiments to indicate the kind of questions that could be answered using field experiments.

Gallino and Moreno (2014) use data from a quasi-experiment that relies on a new 'buy-online, pickup-in-store' functionality being implemented in the US but not in Canada and find that the introduction of 'buy-online, pickup-in-store' leads to a reduction in online sales but an increase in store sales and traffic. Such a study could have presumably be done by randomizing the deployment of a 'buy-online, pickup-in-store' functionality across different US states. Relatedly, Bell et al. (2014) show based on a quasi-experiment that the introduction of an offline channel increases demand overall and through the online channel. Again, it may have been possible to operationalize this as a field experiment, in particular if the 'offline channel' was of a less costly form such as a popup shop.

2.5 Broader Context of Marketing

Last, we address to what extent field experiments are useful when exploring questions of broader importance to marketers. In general, many of the most important questions of marketing strategy, such as whether there is a first-mover advantage, are difficult to analyze using a field experiment technique.

However, recent research suggests that field experiments can be quite useful for analyzing the broader policy or welfare context in which marketing occurs and investigating how marketing can help correct societally charged issues such as inequality in income or across nations. A very useful example of this is the work of Anderson-Macdonald et al. (2015) investigating what parts of a marketing or entrepreneurial education can benefit small startups in South Africa. He finds that in general parts of a curriculum focused on the demand side tended to be more useful than parts of the curriculum focused on the cost side. Another notable feature of this experiment is the mix between digital and non-digital methods in the experimental setting. The educational treatment was done at great expense offline, but data

collected was facilitated and made less costly by the use of digital survey tools to monitor the effects of the treatment.

Digitization and Big Data has also attracted increasing attention to consumer privacy. Miltgen and Tucker (2014) provide some evidence from a field experiment that when money is not involved, people tend to behave in a privacy-protective way which is consistent with their stated privacy preferences. However, when pecuniary rewards are in play, consumers behave inconsistently with their stated privacy preferences, particularly consumers who have the most online experience.[12] A complement to this work on privacy is understanding what makes consumers behave in a non-private way and share information online. Toubia and Stephen (2013) investigate this using a field experiment on Twitter and show that both image-related and intrinsic matter as motivations.

Lastly, field experiments can shed light on a number of broader social issues and serve as real-world validation of laboratory experiments on a variety of topics. Gneezy et al. (2012) examine prosocial behavior in the field and show that initial pro-social acts that come at a cost increase the likelihood of subsequent prosocial acts. Baca-Motes et al. (2013) show that a purely symbolic commitment to an environmentally friendly practice significantly increases this practice. Gneezy and Rustichini (2000) find that the introduction of fines increased late arrivals by parents at day-care centers. Based on a field study in an all-you-can-eat restaurant, Just and Wansink (2011) suggest that individuals are consuming to get their money's worth rather than consuming until their marginal hedonic utility of consumption is zero. Shu et al. (2012) partner with an automobile insurance company and find that signing official documents at the top rather than at the bottom makes ethics more salient and reduces dishonesty. Kivetz et al. (2006) demonstrate in the field that consumption increases

[12]Much work on privacy is limited by firm's unwillingness to experiment with something as legally and ethically sensitive as consumer privacy. Therefore, many papers have taken the approach of Goldfarb and Tucker (2011b); Tucker (2014b) and mixed field experiment data with quasi-experimental changes in privacy regimes.

as consumers approach a reward. Anderson and Simester (2008) use a field experiment that randomized whether there was a surcharge for larger sizes to show that customers respond negatively towards attempts to stigmatize a group by charging a higher price to them.

3 Limitations

Any empirical technique has limitations, and given the special status that field experiments are afforded regarding causal inference in the social sciences, it is particularly important to understand these limitations. We also point our readers to the broader debate in economics about the usefulness of field experiments (see for example Deaton (2009) and Banerjee and Duflo (2008)).

3.1 Lack of Theory

A common critique of field experiments is that they lack theoretical grounding. However, this appears to a be a critique of implementation rather than a critique of method, since a field experiment is purely a statistical technique for obtaining causal inference. It is perfectly viable and indeed desirable for a field experiment to both test and enhance theory. Indeed List (2011) states that 'Experimental results are most generalizable when they are built on tests of [economic] theory'.

One practical way that many field experiments test and enhance theory is by considering different treatment effects in their data, and showing that the treatment effect is larger when theory would predict and absent when theory would predict. Of course one limitation to this approach is that if there is uncertainty about the exact outcome, it is very hard to design field experiments to test a behavioral mechanism at the same time as designing the initial field experiment.

It is worth noting that structural econometric techniques can be combined very well with field experiment data. There is nothing that forces a structural research project to use observational data, and indeed great insights can be gained from the combination of an

economic model and associated modeling with the clarity about the data generating process that is afforded by a field experiment. Examples of researchers who have pursued this path include in economics Duflo et al. (2012) who model dynamic incentives for absenteeism, and in marketing Yao et al. (2012) who use a structural model to evaluate implied discount rates in a field experiment where consumers were randomly switched from a linear to a three-part-tariff pricing plan as well as Dubé et al. (2015) who use two field experiments and a structural model to analyze the role of self-signaling in choices.

Another kind of work in this vein is researchers who use estimates from a field experiment to validate their model. For example, Misra and Nair (2011) used their estimates of differences in dynamic incentives for sales force compensation to implement a field test of new compensation schemes which led to $12 million annually in incremental revenues. Li and Kannan (2014) use a field experiment to evaluate their model for multichannel attribution.

A general challenge with field experiments is clarifying the degree of generalizability of any one study and understanding how the lessons of one point in time will apply in the future.[13] It is perhaps a useful reminder in particular that the aim of a field experiment is not simply to measure a variable at one point in time, but instead to try and measure something that has relevance to both managers and academic theory in the future.

3.2 External Generalizability

An obvious question is how the results of a field experiment conducted, for example in Mexico, will generalize to, say, Norway or India. Without knowledge of the precise primitives that condition a behavioral response among a population, such generalizations are impossible. The same critique would be true of a study based on observational data, and it is important to recognize that a field experiment does not solve this generalizability problem.

Another more subtle critique regarding generalizability is the extent to which the culture

[13]Roberts (1957) urges researchers to ensure that *The underlying conditions of the past persist into the future.*

of the firm that is willing to experiment may affect the results. For example, a firm that is willing to embrace digital experimentation might have other attributes such as superior staff or design interface which aid in unobserved ways the success of the field test. This may potentially limit the generalizability of the findings in other commercial contexts.

Of course one solution to both these challenges is to replicate field experiments across multiple different domains, customers and firms. Such replications allow researchers to understand better the boundaries of any measured effect from in a field experiment context. A good example of the advantages of such an approach is provided by Kremer and Holla (2009) who summarize the learning of several field experiments for the developing world. We also point to Lambrecht et al. (2015), who implement a field experiment with both a charity for homeless people as well as with a fashion firm to confirm their results.

3.3 One-Shot

One practical challenge of field experiments is that they often require substantial effort and/or expense and so a researcher often has only one shot. This has two implications. First, a field experiment 'gone wrong' because of a flaw in the setup, be it theoretical or in the practical implementation, can often not easily be run again, requiring the researcher to carefully consider all possible difficulties and carefully check all practical requirements (e.g., as regards data collection) upfront. Second, it means that researchers can usually implement only a limited set of experimental conditions. As a result, research that aim to demonstrate a more complex behavioral mechanism sometimes complement their field data with laboratory experiments (Berger and Heath, 2008).

3.4 Limited Scope

In the current debate about how appropriate field experiments are for understanding poverty interventions, the director of the World Bank's research department wrote the provocatively entitled 'Should the Randomistas Rule?' (Ravallion, 2009), making the following point:

> From the point of view of development policy-making, the main problem in the randomistas' agenda is that they have put their preferred method ahead of the questions that emerge from our knowledge gaps. Indeed, in some respects (such as the sectoral allocation of research) the randomistas success may have made things worse. The risk is that we end up with lots of social experiments that provide evidence on just one or two parameters for a rather narrow set of assigned interventions and settings. The knowledge gaps persist and even widen.

The same argument could be made within marketing. Field experiment methods are a wonderful way of accurately measuring a causal effect. However, as this article has highlighted, there are some domains of marketing enquiry such as communication and pricing where field experiments are particularly apt, and other areas such as strategy, product development and distribution where field experiment techniques are often more difficult to implement and less likely to be useful. Obviously, this does not mean that such questions should not be asked, but instead that we should be mindful that field experiments have many advantages as a technique but a potentially limited range of applications.

4 In Conclusion

This chapter argues that one of the major advances of the digital age has been to allow digital experimentation. The main advantage of such digital experimentation is to allow causal inference. The challenge now for researchers in this space is to ensure that the causal inferences they are making are both correct given the setting and limitations of any field experiment, and useful in terms of advancing marketing practice.

References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology 69*(2), 334.

Anderson, E. and D. Simester (2003). Effects of $9 price endings on retail sales: Evidence from field experiments. *Quantitative Marketing and Economics 1*(1), 93–110.

Anderson, E. T. and D. I. Simester (2001). Are sale signs less effective when more products have them? *Marketing Science 20*(2), 121–142.

Anderson, E. T. and D. I. Simester (2004). Long-run effects of promotion depth on new versus established customers: Three field studies. *Marketing Science 23*(1), 4–20.

Anderson, E. T. and D. I. Simester (2008). Research note: Does demand fall when customers perceive that prices are unfair? The case of premium pricing for large sizes. *Marketing Science 27*(3), 492–500.

Anderson, E. T. and D. I. Simester (2010). Price stickiness and customer antagonism. *The Quarterly Journal of Economics 125*(2), 729–765.

Anderson-Macdonald, S., R. Chandy, and B. Zia (2015). Returns to business education: The impact of marketing (versus finance) skills on the performance of small firm owners in South Africa.

Andrews, M., X. Luo, Z. Fang, and A. Ghose (2015). Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science*.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Press.

Aral, S. and D. Walker (September 2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science 57*(9), 1623–1639.

Ascarza, E., R. Iyengar, and M. Schleicher (2015). The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. *Journal of Marketing Research*.

Baca-Motes, K., A. Brown, A. Gneezy, E. A. Keenan, and L. D. Nelson (2013). Commitment and behavior change: Evidence from the field. *Journal of Consumer Research 39*(5), 1070–1084.

Banerjee, A. V. and E. Duflo (2008, November). The experimental approach to development economics. Working Paper 14467, National Bureau of Economic Research.

Barrios, T., R. Diamond, G. W. Imbens, and M. Kolesar (2012). Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association 107*(498), 578–591.

Bart, Y., A. T. Stephen, and M. Sarvary (2014). Which products are best suited to mobile advertising? A field study of mobile display advertising effects on consumer attitudes and intentions. *Journal of Marketing Research 51*(3), 270–285.

Bell, D., S. Gallino, and A. Moreno (2014). Showrooms and information provision in omni-channel retail. *Production and Operations Management*.

Berger, J. and C. Heath (2008). Who drives divergence? Identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *Journal of Personality and Social Psychology 95*(3), 593.

Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir, and J. Zinman (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *Quarterly Journal of Economics 125*(1), 263–305.

Blake, T., C. Nosko, and S. Tadelis (2014). Consumer heterogeneity and paid search effectiveness: A large scale field experiment. Technical report, National Bureau of Economic Research.

Boudreau, K. J., N. Lacetera, and K. R. Lakhani (2011). Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science 57*(5), 843–863.

Bruhn, M. and D. McKenzie (2008). In pursuit of balance: Randomization in practice in development field experiments. *World Bank Policy Research Working Paper Series*.

Burtch, G., A. Ghose, and S. Wattal (2015). The hidden cost of accommodating crowdfunder privacy preferences: A randomized field experiment. *Management Science 61*(5), 949–962.

Cook, T. D. and D. T. Campbell (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Houghton Mifflin.

Deaton, A. S. (2009, January). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Working Paper 14690, National Bureau of Economic Research.

Draganska, M., W. R. Hartmann, and G. Stanglein (2014). Internet versus television advertising: A brand-building comparison. *Journal of Marketing Research 51*(5), 578–590.

Dubé, J.-P., X. Luo, and Z. Fang (2015). Self-signaling and pro-social behavior: a cause marketing mobile field experiment.

Duflo, E., R. Hanna, and S. P. Ryan (2012). Incentives work: Getting teachers to come to school. *American Economic Review 102*(4), 1241–78.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica 82*(1), 197–228.

Fong, N. M. (2012). Targeted marketing and customer search. *Available at SSRN 2097495*.

Fong, N. M., Z. Fang, and X. Luo (2015). Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research*.

Gallino, S. and A. Moreno (2014). Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science 60*(6), 1434–1451.

Gneezy, A., U. Gneezy, L. D. Nelson, and A. Brown (2010). Shared social responsibility: A field experiment in pay-what-you-want pricing and charitable giving. *Science 329*(5989), 325–327.

Gneezy, A., U. Gneezy, G. Riener, and L. D. Nelson (2012). Pay-what-you-want, identity, and self-signaling in markets. *Proceedings of the National Academy of Sciences 109*(19), 7236–7240.

Gneezy, A., A. Imas, A. Brown, L. D. Nelson, and M. I. Norton (2012). Paying to be nice: Consistency and costly prosocial behavior. *Management Science 58*(1), 179–187.

Gneezy, U. and A. Rustichini (2000). Fine is a price. *Journal of Legal Studies 29*, 1.

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

Goldfarb, A. and C. Tucker (2014). Standardization and the effectiveness of online advertising. *Forthcoming at Management Science*.

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

Hildebrand, C., G. Häubl, and A. Herrmann (2014). Product customization via starting solutions. *Journal of Marketing Research 51*(6), 707–725.

Hoban, P. R. and R. E. Bucklin (2014). Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research*.

Imai, K., G. King, C. Nall, et al. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science 24*(1), 29–53.

Imai, K., G. King, and E. A. Stuart (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society) 171*(2), 481–502.

Imbens, G. and J. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Jones, S. R. (1992). Was there a Hawthorne effect? *American Journal of Sociology*, 451–468.

Jung, M. H., L. D. Nelson, A. Gneezy, and U. Gneezy (2014). Paying more when paying for others. *Journal of Personality and Social Psychology 107*(3), 414.

Just, D. R. and B. Wansink (2011). The flat-rate pricing paradox: Conflicting effects of 'all-you-can-eat' buffet pricing. *The Review of Economics and Statistics 93*(1), 193–200.

Kim, J.-Y., M. Natter, and M. Spann (2009). Pay what you want: A new participative pricing mechanism. *Journal of Marketing 73*(1), 44–58.

Kivetz, R., O. Urminsky, and Y. Zheng (2006). The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention. *Journal of Marketing Research 43*(1), 39–58.

Kremer, M. and A. Holla (2009). Improving education in the developing world: What have we learned from randomized evaluations? *Annual Review of Economics 1*(1), 513–542.

Lambrecht, A. and C. Tucker (2012). Paying with money or with effort: Pricing when customers anticipate hassle. *Journal of Marketing Research 49*(1), 66–82.

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

Lambrecht, A., C. Tucker, and C. Wiertz (2015). Advertising to early trend propagators? Evidence from Twitter. *Working Paper*.

Lee, L. and D. Ariely (2006). Shopping goals, goal concreteness, and conditional promotions. *Journal of Consumer Research 33*(1), 60–70.

Levav, J., M. Heitmann, A. Herrmann, and S. S. Iyengar (2010). Order in product customization decisions: Evidence from field experiments. *Journal of Political Economy 118*(2), 274–299.

Lewis, R. A. and J. M. Rao (2013). On the near impossibility of measuring the returns to advertising. *Unpublished paper, Google, Inc. and Microsoft Research. http://justinmrao.com/lewis_rao_nearimpossibility.pdf*.

Lewis, R. A. and D. H. Reiley (2014a). Advertising effectively influences older users: How field experiments can improve measurement and targeting. *Review of Industrial Organization 44*(2), 147–159.

Lewis, R. A. and D. H. Reiley (2014b). Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on Yahoo! *Quantitative Marketing and Economics 12*(3), 235–266.

Li, H. A. and P. Kannan (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research 51*(1), 40–56.

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

Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press.

McCarney, R., J. Warner, S. Iliffe, R. van Haselen, M. Griffin, and P. Fisher (2007). The Hawthorne effect: A randomised, controlled trial. *BMC medical research methodology 7*(1), 30.

Meyer, B. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics 12*, 151–162.

Miltgen, C. and C. Tucker (2014). Resolving the privacy paradox: Evidence from a field experiment. *Mimeo, MIT*.

Misra, S. and H. S. Nair (2011). A structural model of sales-force compensation dynamics: Estimation and field implementation. *Quantitative Marketing and Economics 9*(3), 211–257.

Nosko, C. and S. Tadelis (2015). The limits of reputation in platform markets: An empirical analysis and field experiment. Technical report, National Bureau of Economic Research.

Parsons, H. M. (1974). What happened at Hawthorne? New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. *Science 183*(4128), 922–932.

Ravallion, M. (2009). Should the randomistas rule? *The Economists' Voice 6*(2).

Roberts, H. V. (1957). The role of research in marketing management. *Journal of Marketing 22*(1), pp. 21–32.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association 100*(469), 322–331.

Sahni, N. (2011). Effect of temporal spacing between advertising exposures: Evidence from an online field experiment. *Mimeo, Stanford*.

Sahni, N., D. Zou, and P. K. Chintagunta (2014). Effects of targeted promotions: Evidence from field experiments. *Available at SSRN 2530290*.

Schwartz, E. M., E. Bradlow, and P. Fader (2013). Customer acquisition via display advertising using multi-armed bandit experiments. *Ross School of Business Paper* (1217).

Shu, L. L., N. Mazar, F. Gino, D. Ariely, and M. H. Bazerman (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences 109*(38), 15197–15200.

Simester, D., Y. J. Hu, E. Brynjolfsson, and E. T. Anderson (2009). Dynamics of retail advertising: Evidence from a field experiment. *Economic Inquiry 47*(3), 482–499.

Sun, M., X. M. Zhang, and F. Zhu (2012). To belong or to be different? evidence from a large-scale field experiment in china. *NET Institute Working Paper* (12-15).

Tadelis, S. and F. Zettelmeyer (2011). Information disclosure as a matching mechanism: Theory and evidence from a field experiment. *Available at SSRN 1872465*.

Toubia, O. and A. T. Stephen (2013). Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Science 32*(3), 368–392.

Tucker, C. (2014a). Social Advertising. Mimeo, MIT.

Tucker, C. (2014b, October). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research 51*(5), 546–562.

Yao, S., C. F. Mela, J. Chiang, and Y. Chen (2012). Determining consumers' discount rates with field studies. *Journal of Marketing Research 49*(6), 822–841.