On Storks and Babies: Correlation, Causality and Field Experiments

Anja Lambrecht and Catherine E. Tucker

KEYWORDS

Correlation, Causality, Field Experiments, Field Tests, Causal Inference

THE AUTHORS

Anja Lambrecht,
Associate Professor of Marketing,
London Business School, United Kingdom
alambrecht@london.edu

Catherine Tucker,
Professor of Management Science,
MIT Sloan School of Management, Cambridge, USA
cetucker@mit.edu

Correlation is not causality /// The explosion of available data has created much excitement among marketing practitioners about their ability to better understand the impact of marketing investments. Big data allows for detecting patterns and often it seems plausible to interpret them as being causal. While it is quite obvious that storks do not bring babies, marketing relationships are usually less clear. If marketers want to be sure they are not walking into a causality trap, they need to conduct field experiments to detect true causal relationships. In the present digital environment, experiments are easier than ever to undertake, but they need to be prepared and interpreted with great care in order to deliver meaningful and genuinely causal results that help improve marketing decisions.

Apparent causalities often fail to hold up under examination /// The online marketing world is full of examples of organizations or journalists being tempted to make causal inferences from purely correlational data. For example, Twitter on its website reports the information displayed below in Figure 1. In the original headline it stated that engagement with promoted tweets translates to higher brand favorability and purchase intent and suggests that 'this study result highlights the value of an engagement on Twitter.'

In reality, it is difficult to interpret this data as causal. It more likely illustrates that a consumer who views a brand more favorably is also more likely to engage with a promoted tweet by this brand. Similarly, a consumer who intends to purchase a certain brand is more likely to engage with a message promoting this brand. Indeed, the causality could also be reversed. Note that this does not mean the ad is ineffective, but since the data presented is purely correlational it is impossible to judge whether the ad was effective or not.

FIGURE 1:

Correlation interpreted causally:
Does engagement with promoted tweets really translate to higher brand favorability and purchase intent?

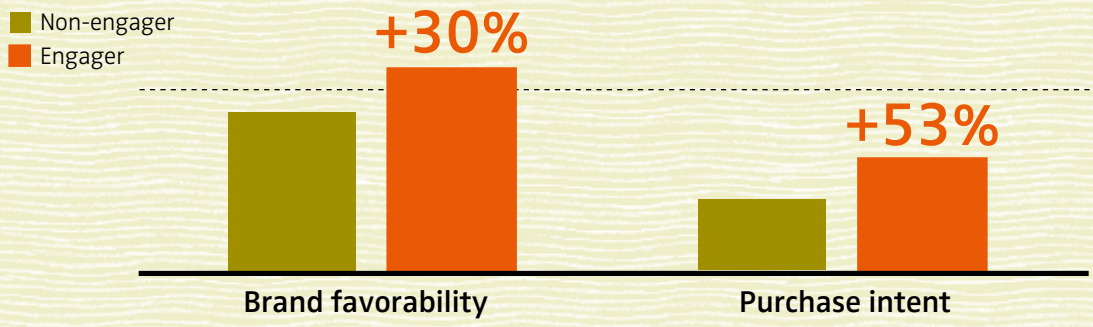
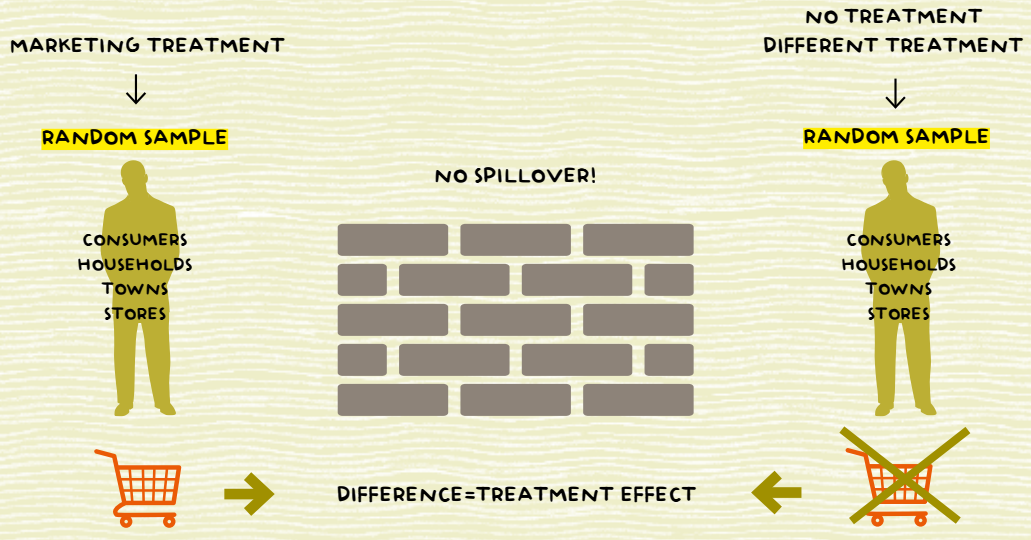


Figure taken from the Twitter website:
https://blog.twitter.com/2013/nielsen-brand-effect-for-twitter-how-promoted-tweets-impact-brand-metrics

FIGURE 2:

How field experiments permit causal inferences



{ *Box 1* }

ONLINE ADVERTISING IS SUCCESSFUL – OR ISN'T IT?

Users who viewed more ads bought more often
Imagine a toy retailer that has implemented a particular form of online advertising: retargeting. Its systems identify users who have looked at its website but did not purchase. As these users continue to browse the web, the toy retailer targets them through ads for the online store. The toy retailer collects detailed user-level data on website visits, ad views, subsequent purchases and non-purchases. The marketing team then evaluates this data. In its analysis it finds that users who viewed more ads were more likely to eventually make a purchase.

Does this mean that the ads were effective in converting users into buyers?
No. The data merely illustrates that users who browsed the web more and, as a consequence, were exposed to more online ads were also more likely to purchase. That's just a correlation. To clarify why such data cannot be interpreted as causal, imagine two users, Emma and Anna. Both Emma and Anna visited the toy retailer's website. In the following weeks, Emma is very busy at work and unable to further attend to Christmas shopping and also unable to browse the internet more broadly. Anna, however, is already on holiday and spends a great deal of time exploring many different gift options online. This means that Emma does not purchase, but merely because she is busy at work and for the same reason she does not look at any online ads. In contrast, Anna has plenty of free time, which leads her to spend a great deal of time on the internet. As a result she is exposed to ads and ultimately buys the product. From the data at hand it is impossible to tell whether Anna's exposure to ads in any way influenced her decision to buy.

But what could the toy retailer do to determine the effectiveness of the ads?
The solution would be a field test as described in Figure 2: It randomly assigns every user who has visited the website to a test group and a control group. The users in the test group will be shown the toy retailer's ads while the users in the control group will be exposed to a replacement ad, such as an ad for a charity. Since, on average, the users in the test group and the users in the control group are the same, any difference in purchase behavior can be attributed to advertising exposure. Reverting to our example, the two groups composed randomly would each include the same number of Annas and Emmas, eliminating the effect of their different behavior.

Field experiments permit causal inferences /// In the social sciences, the gold standard for making causal inference is a field experiment, sometimes referred to as an A/B test. In a field experiment, individual consumers or users are, unbeknown to them, assigned to different groups. One group is then exposed to a marketing treatment, say online advertising, whereas the other group is not exposed to it (see Figure 2).

As long as the company randomly assigns a sufficiently large number of users to each experimental condition, the difference in outcome variable between the two groups of users can be attributed to the marketing treatment. 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 tested effect is hard to predict or assumed to be small. In general, though, it is difficult to give practical advice on sample size beyond aiming for as large a sample and data collection effort as possible. Box 2 highlights the critical decisions necessary to plan and interpret field experiments.

Further applications of field tests to improve marketing decisions /// When the 5 steps described in Box 2 are executed carefully, applications are numerous and we describe some more below.

> ***Comparing the effectivness of generic and personalized ad content*** /// In this study we compared personalized and generic ads for a travel site. Both groups were shown an ad but in one instance users were exposed to a generic brand ad for the site whereas in the other instance the ad

{ *Box 2* }

IMPLEMENTING FIELD EXPERIMENTS SUCCESSFULLY

•

Step 1: Decide on the unit of randomization

Randomization could happen, for example, at the level of the individual, household, town, website, store, or company. While finely-grained units of observation, like single individuals, tend to provide higher statistical power, their setup is often more expensive and difficult to implement. Also, the risk of potential for spillovers and crossovers is higher.

Step 2: Minimize spillovers and crossovers between experimental treatments

Suppose a company randomly selects an individual to receive a free mobile phone. Potentially his or her adoption of a mobile phone could affect the adoption outcomes of relatives and friends even if the relatives and friends were supposedly not treated. If such spillovers are a large concern, 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.

A crossover occurs when an individual who was supposed to be assigned to one treatment is accidentally exposed to another. Suppose, for example, a canned soup company is testing different advertising messages in different cable markets, and individuals are exposed to a different advertising message from that of their home market because they are traveling. This could potentially lead to mismeasurement of the treatment, especially if there were systematic patterns in travel that led to such crossovers not simply being random noise.

Step 3: Decide on complete or stratified randomization

The experimenter then needs to decide whether to conduct stratified 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 more homogenous subsamples. Then each individual in each of these subsets is randomized to a treatment. This stratified technique is useful if some variables are strongly correlated with an outcome. For example, household income may be strongly correlated with purchase behavior toward 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.

Step 4: Ensure that appropriate data is collected

Researchers also need to carefully consider what type of data they need for their later analysis and to ensure that the practical set-up allows them to collect this data. This is especially 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 may require additional steps to ensure that they likewise capture purchase data and can link those to ad exposures.

Step 5: Interpret results from a field experiment carefully

In theory, interpretation of field experimental data should be straightforward, but in practice there are numerous issues to consider when interpreting the statistical results. The key issue is to understand exactly the difference between the groups and to be careful about how to generalize this difference. Also, the duration of the field experiment is critical and will affect the interpretation of results. For example, the researcher needs to have access to a long enough period 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-term effects because experiments are limited in time. Therefore, in most settings researchers should carefully consider whether the causal effect they establish truly reflects the long-term treatment effect.

reflected the specific hotels the user had previously looked at on the company's website. We compared the performance of the different ads and found that on average the generic brand ad was more likely to convert a user to purchase. Only when a consumer's browsing history indicated that they had reached a stage where they were actively comparing attributes of different hotels, did the personalized ads became equally effective.

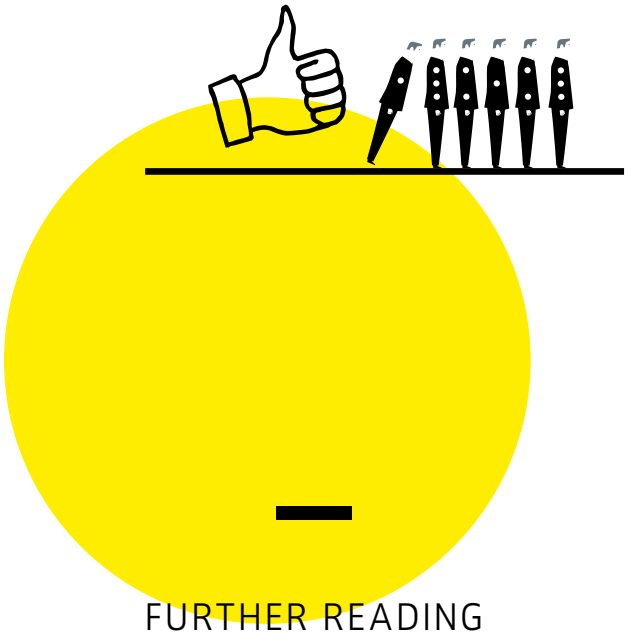
> ***Testing website design*** /// Companies may also wish to compare which of two different designs of their home page is more effective in getting a user to browse products in detail. In this case a company may randomly direct a user to either of the home page versions. The company could then compare the number of users who went on browsing specific products, and later purchased, across the two experimental conditions. Provided that users were randomly assigned to the experimental conditions, the difference in the likelihood to browse or to purchase can be attributed to the difference in the design of the home page.

> ***Optimizing pricing policy*** /// In this article we have mostly focused on marketing communications, but other types of marketing decisions can likewise benefit from insights that come from field experiments. Imagine a company that wishes to estimate how shipping fees affect purchases from their online store. Marketing could set up two different checkout pages where in the first instance the checkout page charges the usual shipping fee and in the second instance the shipping fee is discounted or entirely removed. They could then compare the number of consumers who do not complete their purchase upon reaching the checkout page across conditions and adjust their pricing accordingly.

For companies that want to make sure that they do not invest in storks to get more babies, field experiments represent a very useful avenue in which to obtain truly causal data. When planned and interpreted with care, the results can help to guide a wide range of marketing decisions.

/.

»
In the present digital environment, experiments are easier than ever to undertake.
«



FURTHER READING

Lambrecht, A.; Tucker C. (2015):
"When Does Retargeting Work? Information Specificity in Online Advertising," Journal of Marketing Research, Vol. 50 (5), pp. 561 – 576.

"When Personalized Ads Really Work.,"
https://hbr.org/2013/06/marketers-serve-no-ad-before-i

"Field Experiments in Marketing," working paper.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2630209

Lewis, R. A.; Rao, J. M. (2015):
"The Unfavorable Economics of Measuring the Returns to Advertising," Quarterly Journal of Economics, Vol. 130 (4), pp. 1941 – 1973.