To appear in KDD 2012 Aug 12-16, 2012, Beijing China

# Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained

Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, Ya Xu

Microsoft, One Microsoft Way, Redmond, WA 98052
{ronnyk, alexdeng, brianfra, rogerlon, towalker, yaxu}@microsoft.com

## ABSTRACT

Online controlled experiments are often utilized to make data-driven decisions at Amazon, Microsoft, eBay, Facebook, Google, Yahoo, Zynga, and at many other companies. While the theory of a controlled experiment is simple, and dates back to Sir Ronald A. Fisher's experiments at the Rothamsted Agricultural Experimental Station in England in the 1920s, the deployment and mining of online controlled experiments at scale—thousands of experiments now—has taught us many lessons. These exemplify the proverb that the difference between theory and practice is greater in practice than in theory. We present our learnings as they happened: puzzling outcomes of controlled experiments that we analyzed deeply to understand and explain. Each of these took multiple-person weeks to months to properly analyze and get to the often surprising root cause. The root causes behind these puzzling results are not isolated incidents; these issues generalized to multiple experiments. The heightened awareness should help readers increase the trustworthiness of the results coming out of controlled experiments. At Microsoft's Bing, it is not uncommon to see experiments that impact annual revenue by millions of dollars, thus getting trustworthy results is critical and investing in understanding anomalies has tremendous payoff: reversing a single incorrect decision based on the results of an experiment can fund a whole team of analysts. The topics we cover include: the OEC (Overall Evaluation Criterion), click tracking, effect trends, experiment length and power, and carryover effects.

### Categories and Subject Descriptors

*G.3 Probability and Statistics/Experimental Design:* controlled experiments, randomized experiments, A/B testing.

### General Terms

Measurement, Design, Experimentation

### Keywords

Controlled experiments, A/B testing, search, online experiments

## 1. INTRODUCTION

Online controlled experiments are often utilized to make data-driven decisions at Amazon, Microsoft, eBay, Facebook, Google, Yahoo, Zynga, and at many other companies [1; 2; 3; 4].

Deploying and mining online controlled experiments at large scale—thousands of experiments—at Microsoft has taught us many lessons. Most experiments are simple, but several caused us to step back and evaluate fundamental assumptions. Each of these examples entailed weeks to months of analysis, and the insights are surprising.



We begin with a motivating visual example of a controlled experiment that ran at Microsoft [2]. The team running the MSN Real Estate site (http://realestate.msn.com) wanted to test different designs for the "Find a home" widget. Visitors who click on this widget are sent to partner sites, and Microsoft receives a referral fee. Six different designs of this widget, including the incumbent, were proposed, and are shown in Figure 1.

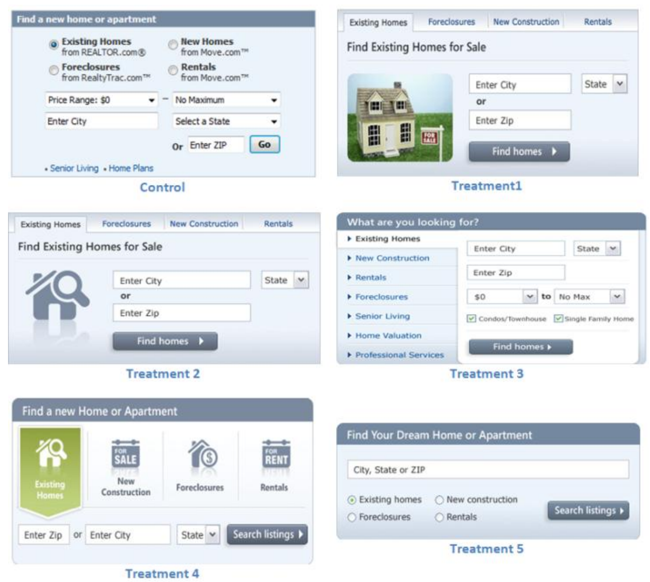



**Figure 1: Widgets tested for MSN Real Estate**

In a controlled experiment, users are randomly split between the variants (e.g., the six designs for the Real Estate widget) in a persistent manner (a user receives the same experience in multiple visits) during the experiment period. Their interactions are instrumented and key metrics computed. In this experiment, the Overall Evaluation Criterion (OEC) was simple: average revenue per user. The winner, Treatment 5, increased revenues by almost 10% (due to increased clickthrough). The Return-On-Investment (ROI) for MSN Real Estate was phenomenal, as this is their main source of revenue, which increased significantly through a simple change.

While the above example is visual, controlled experiments are used heavily not just for visual changes, but also for evaluating backend changes, such as relevance algorithms for Bing, Microsoft's search engine. For example, when a user queries a search engine for "Mahjong," one may ask whether an authoritative site like Wikipedia should show up first, or whether sites providing the game online be shown first. Provided there is agreement on the Overall Evaluation Criterion (OEC) for an experiment, which is usually tied to end-user behavior, ideas can be evaluated objectively with controlled experiments.

One interesting statistic about innovation is how poor we are at assessing the values of our ideas. Features are built because teams believe they are useful, yet we often joke that our job, as the team that builds the experimentation platform, is to tell our clients that their new baby is ugly, as the reality is that most ideas fail to move the metrics they were designed to improve. In the paper *Online Experimentation at Microsoft* [2], we shared the statistic that only one third of ideas tested at Microsoft improved the metric(s) they were designed to improve. For domains that are not well understood, the statistics are much worse. In the recently published book *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society* [5], Jim Manzi wrote that "Google ran approximately 12,000 randomized experiments in 2009, with [only] about 10 percent of these leading to business changes." Avinash Kaushik, author of *Web Analytics: An Hour a Day*, wrote in his Experimentation and Testing primer [6] that "80% of the time you/we are wrong about what a customer wants." In Do It Wrong Quickly [7 p. 240], Mike Moran wrote that Netflix considers 90% of what they try to be wrong. Regis Hadiaris from Quicken Loans wrote that "in the five years I've been running tests, I'm only about as correct in guessing the results as a major league baseball player is in hitting the ball. That's right - I've been doing this for 5 years, and I can only "guess" the outcome of a test about 33% of the time!" [8].

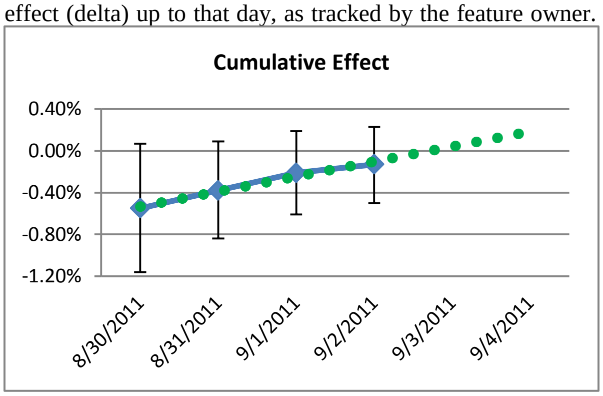
With such statistics, it is critical that the results be trustworthy: incorrect results may cause bad ideas to be deployed or good ideas to be incorrectly ruled out.

To whet the reader's appetite, here is a summary of the five experiments we drill deeper into in this paper, motivated by the surprising findings.

1. Bing, Microsoft's search engine, had a bug in an experiment, which resulted in very poor search results being shown to users. Two key organizational metrics that Bing measures progress by are share and revenue, and both improved significantly: distinct queries per user went up over 10%, and revenue per user went up over 30%! How should Bing evaluate experiments? What is the Overall Evaluation Criterion?

2. A piece of code was added, such that when a user clicked on a search result, JavaScript was executed. This slowed down the user experience slightly, yet the experiment showed that users were clicking more! Why would that be?

3. When an experiment starts, it is followed closely by the feature owners. In many cases, the effect in the first few days seems to be trending up or down. For example, below is the effect from four days of an actual experiment on a key metric, where each point on the graph shows the cumulative effect (delta) up to that day, as tracked by the feature owner.




   The effect shows a strong positive trend over the first four days. The dotted line shows a linear extrapolation, which implies that on the next day, the effect will cross 0% and start to be positive by the sixth day. Are there delayed effects? Primacy effects? Users must be starting to like the feature more and more, right? Wrong! In many cases this is expected and we'll tell you why.

4. From basic statistics, as an experiment runs longer, and as additional users are being admitted into the experiment, the confidence interval (CI) of the mean of a metric and the CI of the effect (percent change in mean) should both be narrower. After all, these confidence intervals are proportional to $1/\sqrt{n}$ when $n$ is the number of users, which is growing. This is usually the case, but for several of our key metrics, the confidence interval of the percent effect does not shrink over time. Running the experiment longer does not provide additional statistical power.

5. An experiment ran and the results were very surprising. This by itself is usually fine, as counterintuitive results help improve our understanding of novel ideas, but metrics unrelated to the change moved in unexpected directions and the effects were highly statistically significant. We reran the experiment, and many of the effects disappeared. This happened often enough that it was not a one-time anomaly and we decided to analyze the reasons more deeply.

Our contribution in this paper is to increase trustworthiness of online experiments by disseminating puzzling outcomes, explaining them, and sharing the insights and mitigations. At Bing, it is not uncommon to see experiments that impact annual revenue by millions of dollars, sometimes tens of millions of dollars. An incorrect decision, either deploying something that appears positive, but is really negative, or deciding not to pursue an idea that appears negative, but is really positive, is detrimental to the business. Anomalies are therefore analyzed deeply because understanding them could have tremendous payoff, especially when it leads to generalized insights for multiple future experiments. The root causes behind these puzzling results are not isolated incidents; these issues generalized to multiple experiments, allowing for quicker diagnosis, mitigation, and better decision-making. Online controlled experimentation is a relatively new discipline and best practices are still emerging. Others who deploy controlled experiments online should be aware of these issues, build the proper safeguards, and consider the root causes mentioned here to improve the quality and trustworthiness of their results and make better data-driven decisions.

The paper is organized as follows. Section 2 provides the background and terminology. Section 3 is the heart of the paper with five subsections, one for each of the puzzling outcomes. We explain the result and discuss insights and mitigations. Section 4 concludes with a summary.

## 2. BACKGROUND and TERMINOLOGY

In the simplest controlled experiment, often referred to as an A/B test, users are randomly exposed to one of two **variants**: Control (A), or Treatment (B) as shown in Figure 2 [9; 10; 11; 12; 3].
There are several primers on running controlled experiments on the web [13; 14; 15; 16]. In this paper, we follow the terminology in *Controlled experiments on the web: survey and practical guide* [17], where additional motivating examples and multiple references to the literature are provided.

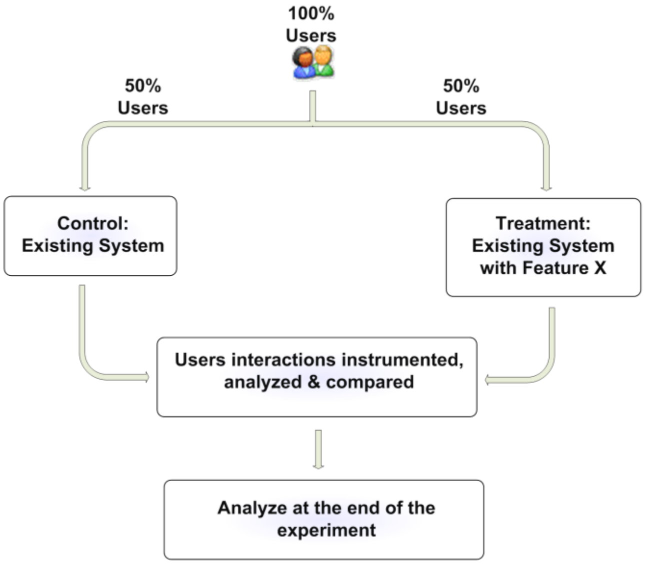



**Figure 2: High-level flow for A/B test**

The **Overall Evaluation Criterion** (OEC) [18] is a quantitative measure of the experiment's objective. In statistics this is often called the **Response** or **Dependent Variable** [9; 10]; other synonyms include **Endpoint**, **Outcome**, **Evaluation metric**, **Performance metric**, **Key Performance Indicator (KPI),** or **Fitness Function** [19]. Experiments may have multiple objectives and a balanced scorecard approach might be taken [20], or selecting a single metric, possibly as a weighted combination of such objectives [18 p. 50].

The **Experimental Unit** is the entity randomly assigned to the control and treatment**.** The examples in this paper use the experimental unit as the analysis unit. For each entity**,** metrics are calculated per unit and averaged over all units in each experiment variant. The units are assumed to be independent. On the web, the user identifier is a common experimental unit, and this is the unit we use throughout our examples.

The **Null Hypothesis**, often referred to as $H_{0,}$ is the hypothesis that the OECs for the variants are not different and that any observed differences during the experiment are due to random fluctuations.

The **Confidence level** is the probability of failing to reject (i.e., retaining) the null hypothesis when it is true. A 95% confidence level is commonly used for evaluating one Treatment versus a Control.

The statistical **Power** is the probability of correctly rejecting the null hypothesis, $H_0$, when it is false. Power measures our ability to detect a difference when it indeed exists.

**Standard Deviation (Std-Dev)** is a measure of variability, typically denoted by $\sigma$.

The **Standard Error (Std-Err)** of a statistic is the standard deviation of the sampling distribution of the sample statistic [9]. For a mean of $n$ independent observations, it is $\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ is the estimated standard deviation.

An experiment effect is **Statistically Significant** if the Overall Evaluation Criterion differs for user groups exposed to Treatment and Control variants according to a statistical test. If the test rejects the null hypothesis that the OECs are not different, then we accept a Treatment as being statistically significantly different from the Control. We will not review the details of statistical tests, as they are described very well in many statistical books [9; 10; 11]. Throughout this paper, statistically significant results are with respect to a 95% confidence interval.

An **A/A Test,** or a Null Test [13] is an experiment where instead of an A/B test, you exercise the experimentation system, assigning users to one of two groups, but expose them to exactly the same experience. An A/A test can be used to (i) collect data and assess its variability for power calculations, and (ii) test the experimentation system (the Null hypothesis should be rejected about 5% of the time when a 95% confidence level is used). The A/A test has been our most useful tool in identifying issues in practical systems. We strongly recommend that every practical system continuously run A/A tests.

# 3. PUZZLING OUTCOMES EXPLAINED

We now review the five puzzling outcomes. These follow the order of the examples in the Introduction. In each subsection, we provide background information, the puzzling outcome, explanations, insights, and ways to mitigate the issue or resolve it.

## 3.1 The OEC for a Search Engine

### *3.1.1 Background*

Picking a good OEC, or Overall Evaluation Criterion, is critical to the overall business endeavor. This is the metric that drives the go/no-go decisions for ideas. In our prior work [12; 17], we emphasized the need to be long-term focused and suggested lifetime value as a guiding principle. Metrics like Daily Active Users (DAU) are now being used by some companies [21]. In *Seven Pitfalls to Avoid when Running Controlled Experiments on the Web* [22], the first pitfall is

> Picking an OEC for which it is easy to beat the control by doing something clearly "wrong" from a business perspective.

When we tried to derive an OEC for Bing, Microsoft's search engine, we looked at the business goals first. There are two top level long-term goals at the President and key executives' level (among other goals): query share and revenue per search. Indeed, many projects were incented to increase these, but this is a great example where short-term and long-term objectives diverge diametrically.

### *3.1.2 Puzzling Outcome*

When Bing had a bug in an experiment, which resulted in very poor results being shown to users, two key organizational metrics improved significantly: distinct queries per user went up over 10%, and revenue per user went up over 30%! How should Bing evaluate experiments? What is the Overall Evaluation Criterion?

Clearly these long-term goals do not align with short-term measurements in experiments. If they did, we would intentionally degrade quality to raise query share and revenue!

### *3.1.3 Explanation*

From a search engine perspective, degraded algorithmic results (the main search engine results shown to users, sometimes referred to as the 10 blue links) force people to issue more queries (increasing queries per user) and click more on ads (increasing revenues). However, these are clearly short-term improvements, similar to raising prices at a retail store: you can increase short-term revenues, but customers will prefer the competition over time, so the average customer lifetime value will decline.

To understand the problem, we decompose query share. Monthly **Query Share** is defined as distinct queries on Bing divided by distinct queries for all search engines over a month, as measured by comScore (distinct means that consecutive duplicate queries by the same user in under half-hour in the same search engine vertical, such as web or images, are counted as one). Since at Bing we can easily measure the numerator (our own distinct queries rather than the overall market), the goal is to increase that component. Distinct queries per month can be decomposed into the product of three terms:

$$\frac{Users}{Month} \times \frac{Sessions}{User} \times \frac{Distinct\ queries}{Session}, \qquad (1)$$

where the $2^{nd}$ and $3^{rd}$ terms in the product are computed over the month, and a session is defined as user activity that begins with a query and ends with 30 minutes of inactivity on the search engine.

If the goal of a search engine is to allow users to find their answer or complete their task quickly, then reducing the distinct queries per task is a clear goal, which conflicts with the business objective of increasing share. Since this metric correlates highly with distinct queries per session (more easily measurable than tasks), we recommend that distinct queries alone not be used as an OEC for search experiments.

Given the decomposition of distinct queries shown in Equation 1, let's look at the three terms

1. Users per month. In a controlled experiment, the number of unique users is going to be determined by the design. For example, in an equal A/B test, the number of users that fall into the two variants will be approximately the same. (If the ratio of users in the variants varies significantly from the design, it's a good indication of a bug.) For that reason, this term cannot be part of the OEC for controlled experiments.
2. Distinct queries per task should be minimized, but it is hard to measure. Distinct queries per session is a surrogate metric that can be used. This is a subtle metric, however, because increasing it may indicate that users have to issue more queries to complete the task, but decreasing it may indicate abandonment. This metric should be minimized subject to the task being successfully completed.
3. Sessions/user is the key metric to optimize (increase) in experiments, as satisfied users will come more. This is a key component of our OEC in Bing. If we had a good way to identify tasks, the decomposition in Equation 1 would be by task, and we would optimize Tasks/user.

Degrading algorithmic results shown on a search engine result page gives users an obviously worse search experience but causes users to click more on ads, whose relative relevance increases, which increases short-term revenue. Revenue per user should likewise not be used as an OEC for search and ad experiments without other constraints. When looking at revenue metrics, we want to increase them without negatively impacting engagements metrics like sessions/user.

### *3.1.4 Lessons Learned*

The decomposition of query volume, the long-term goal for search, reveals conflicting components: some should be increased short term (sessions/user), others (queries/session) could be decreased short term subject to successful task completion. The assumption we make is that a better experience will increase users/month, the last component, which can't be measured in a control experiment.

This analysis is not just impacting search experiments, but also efforts like SEM (Search Engine Marketing). When deciding the bid amount for ads to a search engine, it is natural to try and optimize for the number of queries in the session that started with the ad click. However, long sessions may indicate user frustration (e.g., driving users to mediocre result pages).

Lifetime customer value should typically be the guiding principal for determining your organization's OEC. The choice of specific short-term metrics for controlled experiments needs to be done with a good understanding of the business, and it's critical to understand that long-term goals do not always align with short-term metrics, as shown above.

## 3.2 Click Tracking

### *3.2.1 Background*

Tracking users' online clicks and form submits (e.g., searches) is critical for web analytics, controlled experiments, and business intelligence. Most sites use web beacons (1x1 pixel images requested from a server) to track user actions, but waiting for the beacon to return on clicks and submits slows the next action (e.g., showing search results or the destination page). One possibility is to use a short timeout and common wisdom is that the more time given to the tracking mechanism (suspending the user action), the lower the data loss. Research from Amazon, Google, and Microsoft showed that small delays of a few hundreds of milliseconds have dramatic negative impact on revenue and user experience [17 p. 173], yet we found that many websites allow long delays in order to collect click data reliably. For example, until March 2010, multiple Microsoft sites waited for click beacons to return with a 2-second timeout, introducing an average delay of about 400msec on user clicks. A white paper about the topic was recently published [23]. To the best of our knowledge, this issue is not well understood by most site owners, and implementations have significant click losses. For ads, where clicks are tied to payments, redirects are typically used to avoid click loss. This, however, introduces an additional delay for users and hence not commonly used for tracking clicks.

### *3.2.2 Puzzling Outcome*

A piece of code was added, such that when a user clicked on a search result, additional JavaScript was executed. The reason that piece of JavaScript needed to be executed at that point was that a session-cookie was updated with the destination before the browser was allowed to proceed and open that destination.

This slowed down the user experience slightly, yet the experiment showed that users were clicking more! Why would that be?

### *3.2.3 Explanation*

The "success" of getting users to click more was not real, but rather an instrumentation difference. Chrome, Firefox, and Safari are aggressive about terminating requests on navigation away from the current page and a non-negligible percentage of click-beacons never make it to the server [23]. This is especially true for the Safari browser, where losses are sometimes over 50%. Adding even a small delay gives the beacon more time, and hence more click request beacons reach the server. We have seen multiple experiments where added delays made an experiment look better artificially. Internet Explorer (IE) continues to execute image/beacon requests even after navigation, a decision that relates to backwards compatibility issues, which makes click tracking more reliable.

The above explanation generalizes a previous scenario that we reported [2]. In that scenario, the Hotmail link on the MSN home page was changed to open Hotmail in a separate tab/window. Although the naïve experiment results showed that users clicked more on the Hotmail link when it opened in a new window, the majority of the observed effect was artificial for non-IE browsers. The click was more likely to be logged since opening Hotmail in a separate tab/window did not navigate away from the current page and thus gave the web beacon a greater chance of reaching the server.

In a final example related to click tracking, a change was made in an experiment Treatment for Bing Search so that when a user clicked on a related search, the page would update rather than navigate to a new URL. While the transition looked smoother, we didn't expect such a significant increase to the feature usage as the experiment results showed. In reality, the click logging was simply more reliable and fewer beacon requests were lost. Indeed, the total number of searches (including the related searches) did not increase as much as the clicks had implied.

### *3.2.4 Mitigation*

This problem, although severe, is easy to detect when looking at the experiment results, as Internet Explorer (IE) does not terminate image requests even when navigating away. Thus, if an experiment has an increase in clicks that is attributed to the non-IE browsers, it's likely to be related to the click beacons. This is a good example of a pattern that's easily observable once the underlying root cause is understood. More generally, differences in effects for different browsers are yellow flags for instrumentation issues or differences in HTML/JavaScript parsing.

Long-term, we believe that HTML should have explicit support for beacons.

## 3.3 Initial Effects Appear to Trend

### *3.3.1 Background*

Given the high failure rate for ideas evaluated through controlled experiments as mentioned in the Introduction, it is very common for new ideas to be followed closely in the first few days of the controlled experiment to see if the new idea is a winner or if it should be terminated early.

In this context, it is useful to mention two effects that could occur when new features are introduced: **Primacy** and **Novelty** (or newness) effects [17]. These are opposite effects that sometimes impact experiments. Primacy effect occurs when you change the navigation on a web site, and experienced users may be less efficient until they get used to the new navigation, thus giving an inherent advantage to the Control. Conversely, when a new design or feature is introduced, some users will investigate the new feature, click everywhere, and thus introduce a "novelty" bias that dies quickly if the feature is not truly useful. This bias is sometimes associated with the Hawthorne Effect [24], i.e., a short-lived improvement. The experiment mentioned above, where the Hotmail link on the MSN home page was changed to open Hotmail in a separate tab/window [2] had a strong Novelty effect: users were probably surprised and tried it again several times. While Novelty effects die out after a short duration, and result in a smaller effect, the long-term impact could still be positive, insignificant, or negative. In this case, the long-term effect was positive and the feature is live on the MSN home page.

The existence of Primacy and Novelty effects can be assessed by generating the delta graph (between Control and Treatment) over time, and evaluating trends, visually or analytically. If we suspect such a trend, we can extend the experiment. To evaluate the true effect, an analysis can be done where the OEC is computed only for new users on the different variants, since they are not affected by Primacy and Novelty. Another option is to exclude the first week, as the delta usually stabilizes after a week. But this is where our surprising result comes in: most cases of suspected Primacy and Novelty effects are not real, but just a statistical artifact.

### *3.3.2 Puzzling Outcome*

In many experiments, the effect in the first few days seems to be trending up or down. For example, **Figure 3** shows the effect from the first four days of an actual experiment on a key metric, where each point on the graph shows the cumulative effect (delta) up to that day, as tracked by the feature owner.

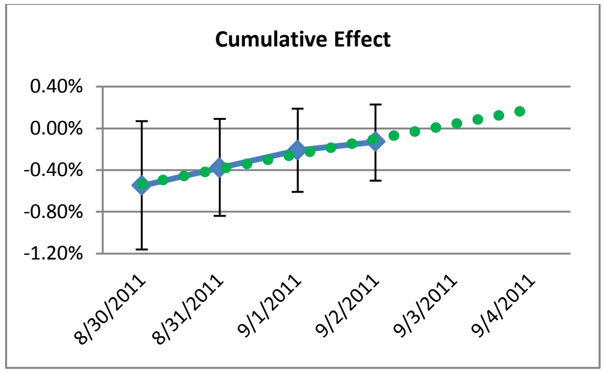



**Figure 3: Effect appears to trend over time**

The effect shows a strong positive trend over the first four days. The experimenter, which is hoping for a positive outcome, sees the initial negative delta, but extrapolates the trend linearly using the dotted line and thinks that on the next day, the effect will cross 0% and start to be positive by the sixth day. The thinking is usually: my feature is obviously great, but it just takes time for users to get used to it, i.e., these are just Primacy effects we're seeing in the first few days. Users must be starting to like the feature more and more, right? Wrong! In many cases this is expected.

### *3.3.3 Explanation*

For many metrics, the standard deviation of the mean is proportional to $1 / \sqrt{n}$, where $n$ is the number of users. For simplicity, assume no repeat users, i.e., each user visits once during the experiment (the results don't change much when using sub-linear growth that happens in practice), so that $n$ is proportional to the number of days. The 95% confidence graph for the measured effect when the actual effect is 0 is shown in Figure 4.

The first few days are highly variable and therefore the effect in the initial days can be much higher or lower than the effect after two or three weeks. For example, the first day has a 67% chance of falling outside the 95% confidence bound at the end of the experiment; the second day has a 55% chance of falling outside this bound. Because the series is auto-correlated, there are two implications

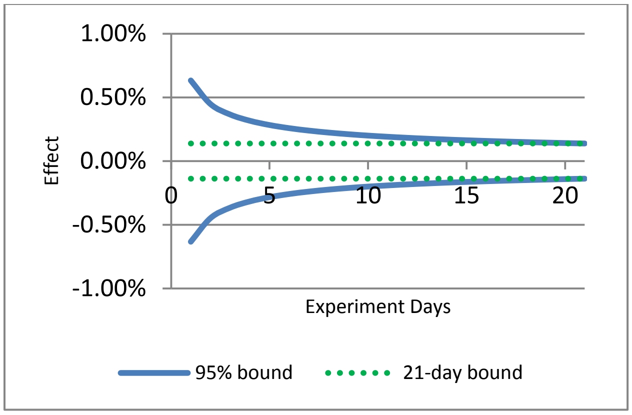



**Figure 4: 95% Confidence Interval over time**

1. The effects in the initial days usually seem overly positive or negative relative to final results published from prior experiments, which ran for much longer. Even if the experiment effect is zero, the initial days show relatively large effects and users typically think: wow, if this 0.8% effect stays, this is a huge win.
2. During the first few days, the cumulative results seem to trend. For example, assume an experiment with no effect on the metric of interest. The first day may be negative at -0.6%, but as more data is accumulated and the effect regresses to the true mean at zero following the 95% confidence cone. Feature owners incorrectly assume the effect is trending and will cross the zero line soon. Of course, this rarely happens.

The graph in Figure 3 actually is from an A/A test (no difference between the control and treatment) where we know the mean of the effect is zero. The first day had a negative delta (note that the wide confidence interval crossed zero) and as more days go by and the confidence interval shrinks, the results regress to the mean. Indeed, as Figure 5 shows, the graph stabilized around zero over time.

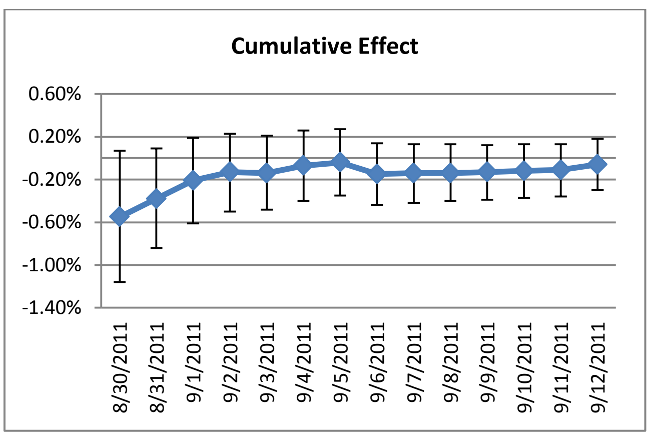



**Figure 5: Effect stabilizes over time**

### *3.3.4 Lessons Learned*

The occurrence of "trends" shown is expected, so we view it as an educational and awareness issue, although we admit that hindsight is 20/20 and we were also fooled by initial trends multiple times. When you've been involved in the implementation of an idea and want it to succeed, the confirmation bias [25] is strong, and the initial negative results are often suppressed as you build a hypothesis that it's trending in the right direction.

Experiments we have run rarely have Primacy effects that reverse the initial effects, i.e., where the feature is initially negative until users learn it and get used to it, then it starts to be positive. We could not find a single experiment where a statistically significant result in one direction became statistically significant in the other direction due to these effects (e.g., a statistically significantly negative becoming statistically significantly positive).

Most experiments have a stable effect (constant mean), but the high variance means that we need to collect enough data to get better estimates; early results are often misleading. While there are true Novelty effects (initially positive effects that die down) or Primacy effects (effects that grows over time), it is more common for a statistically significant negative effect to be more negative over time, and for a statistically significant positive effect to be more positive over time. It is of little value to extend experiments that are statistically significantly negative after a couple of weeks. Failing fast and moving on to the next idea is better.

## 3.4 Experiment Length and Statistical Power

### *3.4.1 Background*

Unlike most offline experiments, online experiments recruit users continuously instead of having a recruitment period before the experiment. As a result, sample size increases as the experiments run longer. One might therefore naturally expect that running an experiment longer provides a better estimate of the treatment effect, and also higher statistical power. Note that for some metrics, such as Sessions/User, the mean increases as the experiment runs longer. We therefore look at percent change relative to the mean in experiments, and likewise calculate power based on percent change.

### *3.4.2 Puzzling Outcome*

For some of our key metrics, including Sessions/user, the confidence interval of the percent effect does not shrink over time. Running the experiment longer does **not** provide additional statistical power for these metrics.

### *3.4.3 Explanation*

The width of the 95% confidence interval of the percentage change can be shown to be determined by two numbers: coefficient of variation and sample size. The coefficient of variation (CV) is the ratio of the standard deviation to the mean, which reflects the level of variation relative to the magnitude of the metric. Assuming Treatment and Control have the same size and same population variance, the width of the 95% confidence interval for percentage change is roughly proportional to:

$$\frac{\text{CV}}{\sqrt{\text{sample size}}}$$

In most statistical applications, samples are modeled as independent and identically distributed, so the CV is determined by the underlying distribution and a larger sample size from the same distribution does not change the CV. Therefore, one naturally expects the 95% confidence interval to shrink by the square root of the sample size as shown in Figure 4. However, empirical data shows that CV does change over time for many online metrics, including Sessions/User. Figure 6 shows the change in the mean, standard deviation and the growth of $\sqrt{\text{sample size}}$ over a 31 day period for a random sample of users. (Note that each series was standardized so they could be plotted on same axis.) The standard deviation is increasing faster than the mean, so CV increases over time.

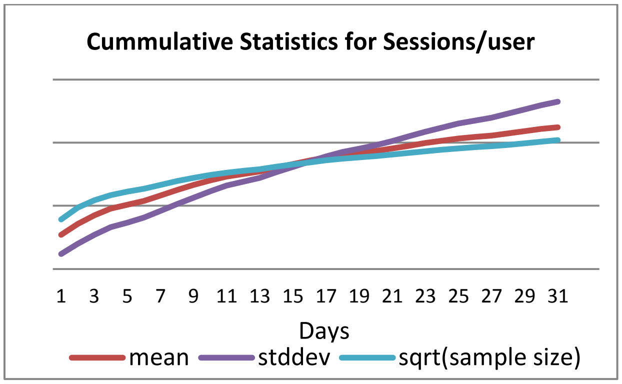



**Figure 6: Change in Mean, Standard Deviation and Sqrt(sample size) for Sessions/user over 31 day period**

In Figure 7, we see that the ratio, CV/√sample-size, is fairly constant (less than 10% change) over the 31 day period.

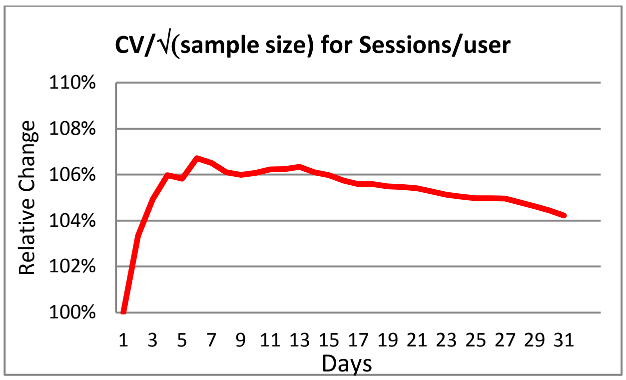



**Figure 7: Change in CV/$\sqrt{\textbf{sample size}}$ for Sessions/User over 31 day period**

The naïve approach is to use the Poisson distribution to model count metrics, such as Sessions/User, but this is clearly a poor model, given our data. The Poisson distribution has mean and variance equal to the parameter $\lambda$. As we run experiment longer, we expect Sessions/User to increase, i.e., the parameter $\lambda$ for the mean is not stationary and increases over time. But even if $\lambda$ were modeled as $\lambda$(t), the CV of Poisson would be $1/\sqrt{\lambda(t)}$, which decreases as $\lambda(t)$ increases, contradicting our empirical data that CV increases over time. Rosset and Borodovsky recently showed that the Negative Binomial is a better way to model count metrics [26]. Moreover, for online user tracking, cookie churn and birth creates further variance for user based metrics. Since this ratio also determines the statistical power, the above result can be rephrased into the following: for metrics like Sessions/user, statistical power does not necessarily increase as an experiment runs longer.

### *3.4.4 Lessons Learned*

For many metrics, especially bounded metrics like clickthrough, the confidence interval for the percent effect shrinks with the experiment duration; running an experiment longer increases statistical power. However, for some metrics like Sessions/user, the confidence interval width does not change much over time. When looking for effects on such metrics, we must run the experiments with more users per day in the Treatment and Control.

If running an experiment longer does not help reduce the width of the confidence interval or increase its statistical power, then why do we run an experiment more than a week? (We consider a week to be the minimum to look at day-of-week effects.) The key reason is that the treatment effect might be delayed due to Primacy and Novelty effects. As we previously noted, this is rare, but we are concerned about the risk.

## 3.5 Carryover Effects

### *3.5.1 Background*

Some online experimentation platforms, including at Bing, Google, and Yahoo, rely on the "bucket system" to assign users to experiments [3]. The bucket system randomizes users into different buckets and then assigns buckets to experiments. It is a flexible system and allows easy reuse of users in subsequent experiments.

### *3.5.2 Puzzling Outcome*

An experiment ran and the results were very surprising. This by itself is usually fine, as counterintuitive results help improve our understanding of novel ideas, but metrics unrelated to the change moved in unexpected directions and the effects were highly statistically significant. We reran the experiment on a larger sample to increase statistical power, and many of the effects disappeared.

### *3.5.3 Explanation*

One big drawback with the "bucket system" is its vulnerability to carryover effects, where the same users who were impacted by the first experiment are being used for the follow-on experiment. This is known, and A/A tests can be run to check for carryover effects, but when they fail, we lose capacity until we re-randomize the bucket assignment. What was surprising to us is the duration of the carryover effect. We share two examples below.

In the first example, we ran the experiment in three stages where we had a 7-day A/A experiment on the user buckets before the A/B experiment was turned on for 47 days. After we finished the experiment, we turned it off and we continued to monitor the same user buckets for more than three weeks. Figure 8 shows the daily percent delta on the OEC (Sessions/User) between the treatment and the control. The gray bars indicate the division for the three stages.

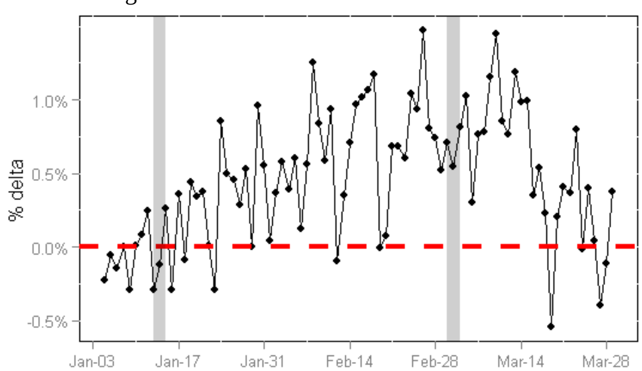



**Figure 8: Carryover Effects Lasted Weeks**

It is clear that there was a carryover effect on users after the experiment finished. The carryover effect seems to die out at about the third week after the experiment.

In another example, shown in Figure 9, a bug exposed users in the experiment to a really bad experience. The carryover effect lasted for much longer. Even after three months, the user buckets still had not fully recovered to their pre-experiment levels.

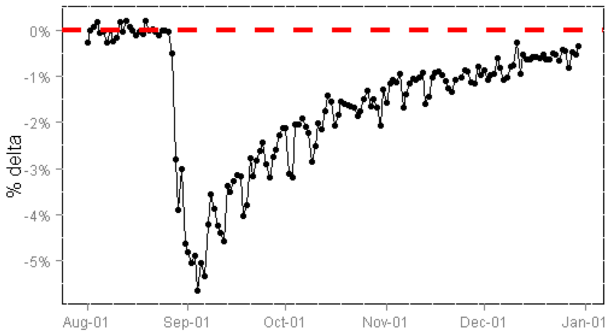
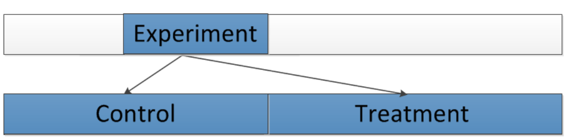



**Figure 9: Long Lasting (3 Months) Carryover Effects**

### *3.5.4 Mitigation*

Even though understanding the carryover effect itself is important to experimenters, it's even more crucial for a quality experimentation platform to guard against it. In the bucket system, because user buckets are recycled from one experiment to another, any carryover impact can easily cause results to be biased for subsequent experiments.

One way to mitigate the issue is local randomization. The root cause of a carryover effect is the fact that the bucket system does not re-randomize for each experiment. The whole bucket system relies on an infrequent bucket re-assignment to randomize the population then the bucket assignment remains constant for a relatively long period. Re-randomizing users into the bucket system can be accomplished by changing the hashing function, but the bucket system couples all experiments in a "bucket line," or "layer" [3], such that we need to stop all running experiments in that bucket line to change the hashing function, hurting capacity and agility. An alternative is to use a two-level bucket system that can accommodate localized re-randomization; that is, re-randomizing only on a subset of buckets, but not affecting others.




The above diagram illustrates how localized re-randomization can be achieved through two-level bucket system. The top level bucket system defines a set of experiment units included in the experiment and treatment assignment resides in the second level bucket system. For each experiment, the second level hashing uses a different hash seed. This guarantees a per-experiment randomization so that treatment assignment is independent of any historical events, including carryover effects from previous experiments. One disadvantage of the above is that one can't use a shared Control: each experiment needs its own Control so that any carryover from an experiment is "mixed" into the Control and Treatment(s).

One nice benefit of localized randomization is that we can run a "retrospective" A/A experiment without actually taking up calendar time. By changing the hashing function, we can re-evaluate the last few days the experiment ran as A/A experiment before starting the A/B experiment. By the independence property of localized re-randomization, if we retrospectively compare users that would have been assigned to control and treatment for any period before the experiment, this comparison will be an A/A. If the A/A shows an effect for key metrics, say p-value < 0.2 (due to an "unlucky" split), we change the hashing key and retry.

## 4. SUMMARY

Controlled experiments are the gold standard in science for proving causality. The FDA, for example, requires controlled experiments (randomized clinical trials) for approving drugs. In the software world, online controlled experiments are being used heavily to make data-driven decisions, especially in areas where the forefront of knowledge is being pushed: Search being a prime example. We are trying a lot of ideas, failing on most, but the successful ones are the ones that help us build a useful theory and apply it. Discovering what works and what doesn't is real Knowledge Discovery in our domain, and it's enabled by a lot of hypothesis generation from mining the data and by running controlled experiments to confirm (or reject) those hypotheses.

The statistical theory of controlled experiments is well understood, but the devil is in the details and the difference between theory and practice is greater in practice than in theory. We have shared five puzzling experiment outcomes, which we were able to analyze deeply and explain. What separates these from many other surprising results is that we found ourselves referring to them over and over. The insight and lessons from the analysis is general and useful not only to other experiments for us, but to many online experiments in other domains, and to many metrics computed for reporting and Business Intelligence reasons.

Generalizing from these puzzles, we see two themes. One is that instrumentation is not as precise as we would like it to be, interacting in subtle ways with experiments: we shared a prime examples where click tracking is brittle. Instrumentation issues are related to the concept of measure validity, i.e., the extent to which the metric used captures the concept we intended to measure [27]. A second theme is that lessons from offline experiments don't always map well online: we have to deal with carryover effects and confidence intervals that don't shrink as we extend the duration and increase the number of users in the experiment. But, fortunately, we get to experiment with very large populations of millions of users and detect small effects to improve our intuition and understanding of the domain.

Anyone can run online controlled experiments and generate numbers with six digits after the decimal point. It's easy to generate p-values and beautiful 3D graphs of trends over time. But the real challenge is in understanding when the results are invalid, not at the sixth decimal place, but before the decimal point, or even at the plus/minus for the percent effect; that's what these analyses did to the initial results. We hope we've managed to shed light on puzzling outcomes and we encourage others to drill deep and share other similar results. Generating numbers is easy; generating numbers you should trust is hard!

## ACKNOWLEDGMENTS

We wish to thank Seth Eliot, Xin Fu, Sara Javanmardi, Greg Linden, David Messner, Saharon Rosset, Maria Stone, Harry Shum, Zijian Zheng, Jing Zhong, and the members of the Experimentation Platform and Bing Data Mining teams.

## References


1. **Kohavi, Ron and Round, Matt.** *Front Line Internet Analytics at Amazon.com.* [ed.] Jim Sterne. Santa Barbara, CA : s.n., 2004. http://ai.stanford.edu/~ronnyk/emetricsAmazon.pdf.

2. **Kohavi, Ron, Crook, Thomas and Longbotham, Roger.** Online Experimentation at Microsoft. *Third Workshop on Data Mining Case Studies and Practice Prize.* 2009. http://exp-platform.com/expMicrosoft.aspx.

3. **Tang, Diane, et al.** Overlapping Experiment Infrastructure: More, Better, Faster Experimentation. *Proceedings 16th Conference on Knowledge Discovery and Data Mining.* 2010, pp. 17-26.

4. **Patil, DJ, et al.** Competing on Analytics at the Highest Level. [Online] Apr 9, 2010. http://www.svforum.org/index.cfm?fuseaction=Page.ViewPage&PageID=997.

5. **Manzi, Jim.** *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society.* s.l. : Basic Books, 2012. 978-0-465-02931-0.

6. **Kaushik, Avinash.** Experimentation and Testing: A Primer. *Occam's Razor.* [Online] May 22, 2006. http://www.kaushik.net/avinash/2006/05/experimentation-and-testing-a-primer.html.

7. **Moran, Mike.** *Do It Wrong Quickly: How the Web Changes the Old Marketing Rules* . s.l. : IBM Press, 2007. 0132255960.

8. —. Multivariate Testing in Action: Quicken Loan's Regis Hadiaris on multivariate testing. *Biznology Blog by Mike Moran.* [Online] December 23, 2008. http://www.biznology.com/2008/12/multivariate_testing_in_action/.

9. **Mason, Robert L, Gunst, Richard F and Hess, James L.** *Statistical Design and Analysis of Experiments With Applications to Engineering and Science.* s.l. : John Wiley & Sons, 1989. 047185364X .

10. **Box, George E.P., Hunter, J Stuart and Hunter, William G.** *Statistics for Experimenters: Design, Innovation, and Discovery.* 2nd. s.l. : John Wiley & Sons, Inc, 2005. 0471718130.

11. **Keppel, Geoffrey, Saufley, William H and Tokunaga, Howard.** *Introduction to Design and Analysis.* 2nd. s.l. : W.H. Freeman and Company, 1992.

12. **Kohavi, Ron, Henne, Randal M and Sommerfield, Dan.** Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. *The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007).* August 2007, pp. 959-967. http://exp-platform.com/hippo.aspx.

13. **Peterson, Eric T.** *Web Analytics Demystified: A Marketer's Guide to Understanding How Your Web Site Affects Your Business.* s.l. : Celilo Group Media and CafePress, 2004. 0974358428.

14. **Eisenberg, Bryan.** How to Improve A/B Testing. *ClickZ Network.* [Online] April 29, 2005. http://www.clickz.com/clickz/column/1717234/how-improve-a-b-testing.

15. **Chatham, Bob, Temkin, Bruce D and Amato, Michelle.** *A Primer on A/B Testing.* s.l. : Forrester Research, 2004.

16. **Miller, Scott.** How to Design a Split Test. *Web Marketing Today, Conversion/Testing.* [Online] Jan 18, 2007. http://www.wilsonweb.com/conversion/.

17. **Kohavi, Ron, et al.** Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery.* February 2009, Vol. 18, 1, pp. 140-181. http://exp-platform.com/hippo_long.aspx.

18. **Roy, Ranjit K.** *Design of Experiments using the Taguchi Approach : 16 Steps to Product and Process Improvement.* s.l. : John Wiley & Sons, Inc, 2001. 0-471-36101-1.

19. **Quarto-vonTivadar, John.** AB Testing: Too Little, Too Soon. *Future Now.* [Online] 2006. http://www.futurenowinc.com/abtesting.pdf.

20. **Kaplan, Robert S and Norton, David P.** *The Balanced Scorecard: Translating Strategy into Action.* s.l. : Harvard Business School Press, 1996. 0875846513.

21. **Cheshire, Tom.** Test. Test. Test: How wooga turned the games business into a science. *Wired Magazine UK, http://www.wired.co.uk/magazine/archive/2012/01/features/test-test-test?page=all.* January 5, 2012.

22. **Crook, Thomas, et al.** Seven Pitfalls to Avoid when Running Controlled Experiments on the Web. [ed.] Peter Flach and Mohammed Zaki. *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2009, pp. 1105-1114. http://exp-platform.com/ExPpitfalls.aspx.

23. **Kohavi, Ron, et al.** *Tracking Users' Clicks and Submits: Tradeoffs between User Experience and Data Loss.* Redmond, WA : s.n., 2010. Microsoft White pape, http://www.exp-platform.com/Pages/TrackingClicksSubmits.aspx.

24. Hawthorne effect. *Wikipedia.* [Online] 2007. http://en.wikipedia.org/wiki/Hawthorne_experiments.

25. **Nickerson, Raymond S.** Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology.* 1998, Vol. 2, 2, pp. 175-220.

26. **Rosset, Saharon and Borodovsky, Slava.** A/B Testing Using the Negative Binomial Distribution in an Internet Search Application, Submitted. [Online] 2012. http://www.tau.ac.il/~saharon/papers/AB%20testing%20with%20NB%20distribution%20-%20revision.pdf.

27. **Weiss, Carol H.** *Evaluation: Methods for Studying Programs and Policies.* 2nd. s.l. : Prentice Hall, 1997. 0-13-309725-0.