Unexpected Results in Online Controlled Experiments

Ron Kohavi
Microsoft
1 Microsoft Way
Redmond, WA 98052
ronnyk@microsoft.com

Roger Longbotham
Microsoft
1 Microsoft Way
Redmond, WA 98052
rogerlon@microsoft.com

ABSTRACT

Controlled experiments, also called randomized experiments and A/B tests, have had a profound influence on multiple fields, including medicine, agriculture, manufacturing, and advertising. Offline controlled experiments have been well studied and documented since Sir Ronald A. Fisher led the development of statistical experimental design while working at the Rothamsted Agricultural Experimental Station in England in the 1920s. With the growth of the world-wide-web and web services, online controlled experiments are being used frequently, utilizing software capabilities like ramp-up (exposure control) and running experiments on large server farms with millions of users. We share several real examples of unexpected results and lessons learned.

Keywords

Controlled Experiments, A/B tests, statistics, unexpected results.

1. INTRODUCTION

Any figure that looks interesting
or different is usually wrong

-- Twyman's law

In the online world, controlled experiments allow the evaluation of ideas by exposing users to different variants. It is said that ideas are like children: everyone likes their own; however, our experience is such that most ideas, even those that pass all organizational bars and get implemented, fail to improve the metrics they were designed to improve (1). A survey of controlled experiments and a practical guide is available elsewhere (2) and excellent books on the topics of experiments exist (3; 4; 5). In this paper, we do not review specific ideas (even though many are unexpected) but rather we share unexpected results related to the proper execution of controlled experiments. One of the simplest designs for a controlled experiment is called an A/B test, where users are randomly assigned to either the standard, or default, site known as the Control or version A, and the remaining users are assigned to the Treatment, or version B, containing changes to test.

One of the most important recommendations we have for anyone running online controlled experiments is to run A/A tests (2; 6). An A/A test is similar to an A/B test in that the software exercises the user split, but both populations are shown the same experience. Observations are collected, metrics are computed, and the A/A test should show no statistically significant difference 95% of the time (if 95% confidence intervals are used). Having run many A/A tests, we have seen many unexpected results that provided us with appreciation for how the slightest differences could result in significant changes to the user experience. We share multiple examples of failed A/A tests.

We share ten examples of unexpected results; we explain the reasons (often the results of very expensive investigations), and share the lessons. Anomalies are expensive to investigate, but we found that some lead to critical insights that have long-term impact. We hope we can save you, the reader, investigation time by sharing our insights and lessons.

2. BROWSER REDIRECTS

A very common and practical mechanism used to implement an A/B test is to redirect the treatment to another page. Like many ideas, it is simple, elegant, and wrong; several different attempts have shown that this fails an A/A test (or rather the A/A' test, where A' uses a redirect). The implementation is as follows: if the randomization function determines that the user should be in Control, the page is displayed; if the randomization shows that the user should be in Treatment, a browser redirect is done by using the http-equiv="REFRESH" meta tag in HTML. In every case where we have conducted this as an A/A' test the version with the redirect significantly underperformed the other version. The reasons for this unexpected difference are:

1. Performance differences. Users in the Treatment group suffer an extra redirect, which may appear fast in the lab, but delays for users may be significant, on the order of hundreds of milliseconds. Slowdowns on this scale have significant impact on metrics. See, for example, Speed Matters in the survey paper (2).
2. Bots. Different robots will handle redirects differently: some may not redirect, some will tag this as a new page worthy of deep crawling, etc. As long as bots are distributed uniformly in the Control and Treatment, their relative impact is small. However, in this case subtle biases are being introduced, causing the A/A tests to fail, indicating that an A/B test will be biased.
3. Redirects are asymmetric. When users are redirected to the treatment page, they may bookmark it or pass a link to their friends. Bots might add this new page to their index for crawling. In most implementations, the Treatment page does not check that the user should really have been randomized into the Treatment and hence there is contamination.

The lesson here, first noted in (6) is to avoid redirects in implementations and prefer a server-side mechanism that generates HTML. When that is not possible, make sure that both Control and Treatment have the same "penalty." We sometimes run an A/A'/B' test, where the A' and B' are redirected. The comparison between A' and B' is therefore fair, and the difference between A and A' gives us an idea of the impact of the redirect on key metrics.

3. EXPOSURE CONTROL

In several situations, we saw surprising results that were traced to bad exposure control, i.e., which users are exposed to the experiment variants. Most of these are obvious in hindsight, but raising awareness of the issue up front may save significant time. Some examples

1. The MSN US Home Page redirects users from some countries to their local country: if you visit www.msn.com from an IP in India or the UK, the assumption is that you want to see the local MSN Home Page and are thus redirected automatically or semi-automatically (a popup shows up with a question). Many international sites (e.g., Google) implement this reverse-IP lookup to raise awareness of their local sites and help users. When a new version of the MSN US Home Page was tested in a controlled experiment, the reverse-IP lookup was not yet implemented for the new page. The results were highly biased because the population of users from non-US IPs was much higher in the Treatment than in the Control.
2. In a Bing experiment, a misconfiguration caused all Microsoft users to always see Control. This created enough of a bias to skew results.

The lesson here is to run A/A tests that resemble the final setup as close as possible and also to drill down and slice the data by common attributes, such as country and browser. Large differences may hint at improper exposure control.

4. SHARED RESOURCES

When running controlled experiments with two variants, the highest overall power is achieved when the population split is 50%/50%. In practice, treatments may need to run at lower percentages. For example, during ramp-up of an experiment, one should start at very lower percentages; for very large sites, there may be enough power with a small percentage of users; if multiple disjoint experiments need to be run, they may share a control, which would be larger; if one is interested in running comparisons between different treatments to the Control, a larger control provides more power (2).

We usually run A/A tests at 50%/50%, but we were surprised when a 90%/10% A/A test failed consistently. It turns out that a bounded resource is the cause. In this case an LRU (least-recently-used) cache was used, and the entries for the Control and Treatment were disjoint. Because the experiment ran as a 90%/10% experiment, the Control had significantly more entries in the LRU cache, leading to a higher cache-hit ratio and thus better performance, impacting the user experience and leading to better metrics for the control (6).

The lesson here is to be aware of possible issues with shared resources. As always, start with A/A tests and be vigilant about measuring performance.

5. BROWSER DIFFERENCES

The MSN home pages have a link to Hotmail, which is heavily used. In the UK, we tested whether the link should open Hotmail in a new window rather than in place. As we reported (1), engagement increased significantly and despite some concerns about the "pop-up" this was deployed. We repeated the experiment in the US and looked at additional metrics. One metric that was statistically significantly higher in Treatment than in Control and raised a red flag was the percentage of users who clicked on the Hotmail link (an indicator variable). Clicking on this link is the triggering point (2), so there should not be a statistically significant difference until *after* the users click, as this is the first point where something differs (a new window is opened for the Treatment group). With such an unexpected result, we sliced the data by multiple variables and type of browser used had highly significant variations in the Treatment effect. (The reason for the difference is that clicks are commonly instrumented using a web beacon or web bug (7), a small 1x1 image being requested from the server asynchronously using JavaScript, but the mechanism is well known to be lossy, i.e., not every click beacon makes it to the destination server. The reliability of the beacons could be increased by waiting for the beacon, but most sites choose to wait a fixed time and not slow the user experience, resulting in some loss of clicks.) In this case, it turns out that by opening the destination in a new window, the beacon's reliability improved significantly for non-IE browsers and hence the delta. The value of the feature was still positive once we corrected for the instrumentation issue, but not as high as the initial results.

There are multiple lessons here:

1. Investigate anomalies seriously. In this case, the indicator variable (percent of users using Hotmail) was unexpected. After the instrumentation correction, it was statistically insignificant, as expected.
2. Proactively drill-down by key attributes, such as browser family (based on user-agent) and geography (based on reverse IP)

6. LONG-TERM OEC

When the first author joined Amazon, there were campaigns that sent e-mails to users, introducing them to products they may be interested in. Here are snippets from recent e-mails explaining the concept:

1. As someone who has purchased Xbox 360 consoles or games at Amazon.com, you might like to know that you can play Kinect for Xbox 360 on day one with Release-Date Delivery
2. As someone who has browsed or purchased Wii products at Amazon.com…
3. As someone who has shown an interest in Mrs. May's snacks…, you might like to know about the following offer…
4. Customers who purchased books on entrepreneurship from Amazon.com…

Given an event, a product, or a family of products to promote, there are many ways to pick a population of users who purchased or browsed other products. The evaluation of ideas for these campaigns was done using controlled experiments, with some targeted users being excluded from the e-mail and serving as the Control, a standard industry practice. The OEC (Overall Evaluation Criterion) for the campaigns was based on purchases whose sessions were referred by the e-mails. But under this criterion, all ideas/campaigns evaluated positively, causing a large number of e-mails to be sent and customers complained. Mechanisms were introduced to limit the frequency of e-mails to users, but this was the wrong approach, as the OEC was not taking into account the negative impact of "spam."

The OEC was then refined to look at long-term customer value and a campaign (or campaign family) was penalized for unsubscribes, as these customers are no longer targetable in future campaigns. Once the penalty based on the number of unsubscribes times their lifetime value from e-mail was taken into account, many campaigns evaluated negatively, a result that surprised many people, but that users loved. Campaigns had to be better targeted and with higher value for users to pass the new higher OEC bar.

The lesson here is obvious in hindsight: pick the OEC carefully and try to model the customer lifetime value, not short-term benefits (2).

7. MONITORING SYSTEMS

We tested a new design for a page shown to users who run a non-genuine version of Windows, prompting them to buy a valid product key. The specifics are not important, as the example applies to any online retail site with a checkout/purchase button. The OEC (Overall Evaluation Criterion) was simple: of the users who see the page, what percent click the "buy" button to initiate a purchase, a classical one-step conversion metric. The new page had a much lower conversion rate, but one surprising anomaly was that the number of page views per user was significantly up for the Treatment. An investigation revealed that the experimenting site had a monitoring system that requested the page and then simulated a click on the purchase button and checked the ordering pipeline. The system was designed such that if the click failed, it would try multiple times before raising an alarm. It turned out that with the new Treatment design, the "click" action from the monitoring system did not work and it made many retries, reducing the click-through rate for the Treatment.

The lesson here is to take bots and monitoring systems into account. We have previously discussed the impact of robots (8). Monitoring systems can create large skews if unaddressed.

8. UNPLANNED DIFFERENCE BETWEEN VARIANTS

We were conducting an experiment comparing the ordering of headlines on the MSN Home Page (8). The Control was editorially driven and the Treatment was a randomized order. We expected some degradation in engagement, as measured by clicks. When we analyzed the results of the experiment, we found the randomly placed headlines had a 2% increase in clicks and was highly significant (p-value<0.001).

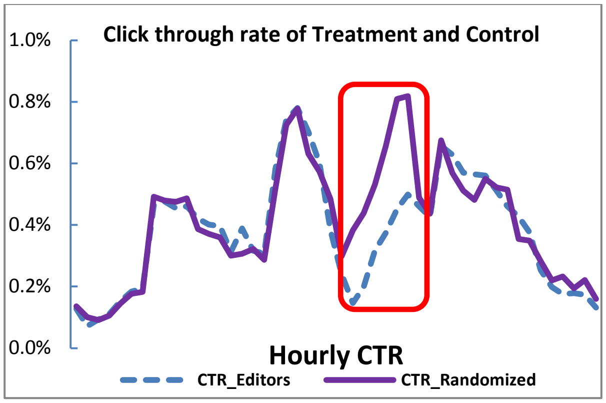

Figure 1 Click through rate showing 7 hour period with unplanned difference

An investigation started, where we drilled down to hourly data. A plot of the hourly click-through rate (CTR) showed a seven hour period where the randomized group performed better (Figure 1). Otherwise the two groups looked about the same.

We investigated what could be causing this difference and found that the top headline for the two groups referred to different stories for this seven hour period, an uncontrolled difference.

Several lessons are important to mention here:

1. Experimental control is critical. Keep everything constant except the thing you want to test.
2. Drill-down by time to look at hourly data. Had the result not been so surprising (e.g., if the treatment were 2% worse), we might have accepted the result. We now regularly show hourly plots for sanity checks to detect such anomalies.
3. Use screen scrapers to save screen shots of the pages being experimented on a regular basis in order to allow debugging of surprises. We have found this to be extremely useful in other experiments.

9. SIMPSON'S PARADOX

One experiment showed the Treatment was 4% worse than the Control. We plotted the effect by day and saw the Treatment was better than the Control on almost every day (Figure 2). What's going on?

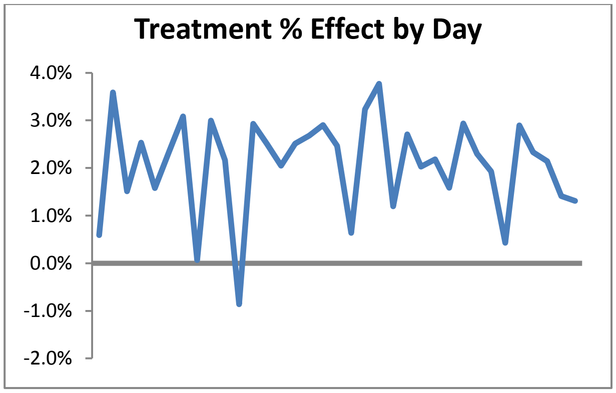

Figure 2 Daily Treatment effect for experiment with overall -4% Treatment effect

One feature of this experiment is that it was a ramp-up, meaning that the percentage of users in the Treatment increased during the experiment. An experimenter may want to do this if the Treatment has a risk of a large negative effect, e.g. due to bugs or adverse customer reaction.

Figure 3 shows the means for the two groups as well as the percentage of users in the Treatment. The experiment ran for five weeks, starting on a Monday. The Treatment had 1% of users for the first 26 days when it went to 5% of users for one week then to 50% of users for the last two days. This metric follows the usual pattern of clicks per user being lower on the weekends than weekdays, so the last two days had fewer clicks per user than the average, but the Treatment effect was still positive. However, since the Treatment had a much larger percentage on the last two days, the clicks per user on those days carried larger weight with the Treatment mean making the Treatment look worse than the Control. This is a good example of Simpson's paradox (9; 10; 11). There are special analyses options you can take to make sure Simpson's paradox doesn't impact your results, or you can simply require that the percentage of users in the Treatment relative to the Control not change during the experiment. In the latter case, any ramp-up period must be completed prior to the start of the experiment.

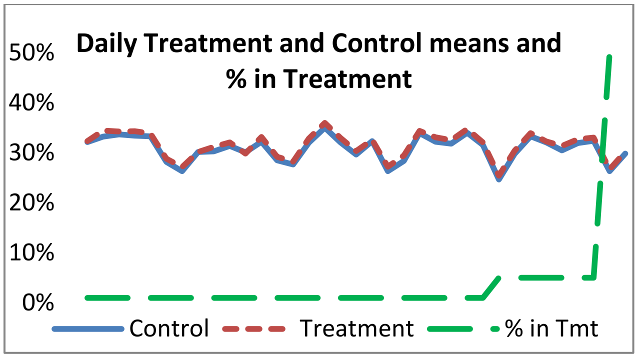

Figure 3 Treatment and Control means and ramp-up percentage for Treatment

Lesson: Beware the potential impact of Simpson's paradox

10. TWO MORE UNEXPECTED RESULTS

These two unexpected results were shared in another paper (8) but they are so important we wanted to reference them here. These phenomena can have a large impact and can affect any experiment. For the first, Office Online was testing a redesigned homepage that looked more modern, was a cleaner design and had fewer links to distract from the primary objective, getting users to click on the buttons to download a version of Microsoft Office for trial or purchase. The primary objective was increasing the number of downloads. Figure 4 shows the old homepage on the top and the newer version below. The red squares outline the areas where a user clicks to take them to the download center where they either purchase or download a trial version of office.

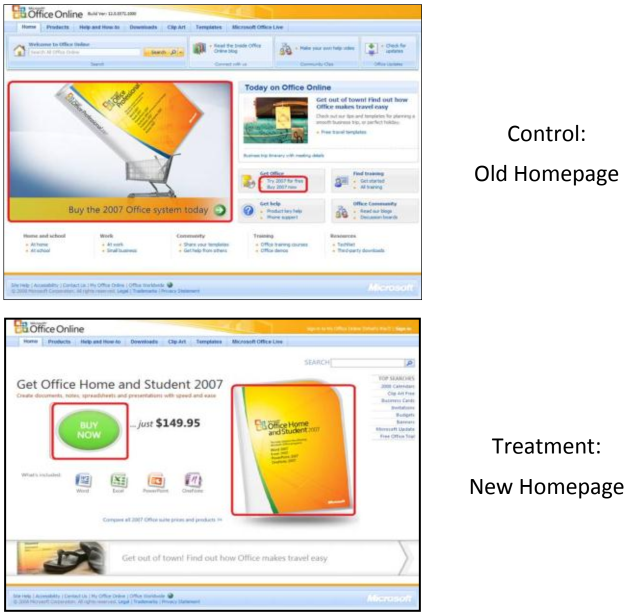
Office Online
Today on Office Online
Buy the 2007 Office system today
Control:
Old Homepage
Office Online
Get Office Home and Student 2007
BUY NOW
... just $149.95
Get out of town! Find out how Office makes travel easy
Treatment:
New Homepage

Figure 4 Old and New Designs for Office Online Homepage

Instead of the number of clicks to the download buttons going up as expected, they decreased 64%! When such a large unexplained delta is seen, one should look for a mistake in the experiment or the assumptions. Upon examination of the design the words in the Treatment button are "Buy Now" with the $149.95 price, whereas the words in the small corresponding link in the Control are "Try 2007 for free" and "Buy now." So, even though the design may be better in the Treatment, it is well known that the offer of something for free has a huge psychological advantage **(12)**. In addition, the Treatment shows the price of this version of Office whereas the Control does not give the price. It is well known that product pages such as the Control where the price is not shown will have many more clicks to "add to cart" to get more information, namely the price. This does not mean there are more purchases, but rather that the conversion rate during the purchase pipeline may be different with the Treatment sending more qualified users to the pipeline.

Lesson: Always get information on the ultimate action you want the user to take.

For the second of these surprises we took a real experiment and simulated an A/A experiment by rerandomizing users into the two groups and doing the calculation of treatment effect for all metrics. We did this 6,000 times. One set of metrics had 5% statistically significant, which was the expected Type I error rate. However, another type of metric was statistically significant 30% of the time. The reason for this was the way in which we calculated standard deviation. In both cases we used the standard statistical formula for standard deviation but the metrics that had 5% significant had uncorrelated experimental units. The second type of metric had positively correlated units which gave an underestimate of standard deviation by two-thirds. We now use bootstrapped estimates of standard deviation for the latter type of metrics (13).

Lesson: Beware of classical statistical formulas that assume independence.

11. CONCLUSION

We have given many examples where unexpected and incorrect results were seen in online randomized experiments. Almost all of these are due to subtle errors that are not easy to anticipate or detect unless the experimenter is looking for them. We recommend an online experimenter make frequent use of A/A experiments, segment the results by key attributes such as browser and conduct data quality checks that can detect some of the more frequent problems. If a result seems unexpected it may be due to lack of understanding of user behavior or it could be due to a software or experimental design problem. You want to be able to rule out the latter if at all possible. Paraphrasing Twyman's law, if a result is truly unexpected, it's probably wrong. Of course, it's not always true, but we have learned it pays to be skeptical of results are surprising.

Finally, there is one meta-lesson we have learned from running many online experiments: "Getting numbers is easy, getting numbers you can trust is quite difficult." Running a good online experiment is a lot more than just randomly assigning users into two groups – it requires careful planning and vigilance in monitoring for known and yet-to-be discovered sources of experimental bias.

12. ACKNOWLEDGMENTS

We would like to thank members of the Experimentation Platform team at Microsoft. Special thanks to David Treadwell., Ray Ozzie, and Eric Rudder; without their support the experimentation platform would not have existed.

13. Bibliography

1. **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.

2. **Kohavi, Ron, et al., 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.

3. **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.

4. **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 .

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

6. **Kohavi, Ron, Longbotham, Roger and Walker, Toby.** Online Experiments: Practical Lessons. [ed.] Simon S.Y. Shim. *IEEE Computer.* September 2010, Vol. 43, 9, pp. 82-85. http://exp-platform.com/IEEE2010ExP.aspx.

7. **Wikipedia.** *Web Bug.* [Online] 2010. http://en.wikipedia.org/wiki/Web_bug.

8. **Thomas Crook, Brian Frasca, Ron Kohavi, Roger Longbotham.** 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.

9. **Simpson, Edward H.** The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, Ser. B.* 1951, Vol. 13, pp. 238–241.

10. **Malinas, Gary and Bigelow, John.** Simpson's Paradox. *Stanford Encyclopedia of Philosophy.* [Online] 2004. [Cited: February 28, 2008.] http://plato.stanford.edu/entries/paradox-simpson/.

11. **Wikipedia: Simpson's Paradox.** Simpson's paradox. *Wikipedia.* [Online] 2008. [Cited: February 28, 2008.] http://en.wikipedia.org/wiki/Simpson%27s_paradox.

12. **Ariely, Dan.** *Predictably Irrational.* New York : HarperCollins Publishers, 2009.

13. **Efron, Bradley and Robert J. Tibshirani.** *An Introduction to the Bootstrap.* New York : Chapman & Hall, 1993. 0-412-04231-2.

About the authors

Ron Kohavi is the general manager of Microsoft's Experimentation Platform (ExP) and Roger Longbotham is the team's analytics manager. For more information about ExP, see http://exp-platform.com/.