Basics of A/B Experimentation, Part 2

Roger Longbotham Process Performance Management https://ppmdatascience.solutions

The second White Paper in the series where I am giving the basic knowledge one needs to successfully conduct A/B experiments online.

In this White Paper I will be addressing the following topics:

- More on metrics (Introduction to metrics in Part 1)
- Presentation of results
- Randomization and analysis units
- Analysis of ratios
- Initial experiment diagnostic checks
- A/A tests

More on Metrics

In Part 1 of this series I gave an introduction to choosing metrics for online experimentation. A major emphasis is for the experimenter to choose one Primary Metric (PM) that she/he expects to be impacted by the experiment. A large component of the decision as to whether to implement the Treatment is whether the PM is statistically significant and positive. Of course, there are other considerations before you decide to turn on the Treatment to replace the Control experience. Some of those considerations:

- Are other key metrics impacted?
- Is this a short-term effect?
- Are there any indications the result is not trustworthy?

The issue of the effect seen during the experiment possibly being a short-term effect will be dealt with in Part 3 of this series. [We are assuming other questions such as

- Is this change consistent with the organization's goals and direction for the site? and
- Are there any ethical issues with making this change? (See Part 3)

have already been resolved prior to starting the experiment.]

Metric Hierarchy

It is important that an organization define one metric (or a small set of metrics) to reflect the overarching goal of the site. This is often called the **Overall Evaluation Criterion (OEC).** You may think of this as the "north star" for the site. Any change that degrades the OEC should not be implemented. For example, the OEC for an ecommerce site may be a combination of customer loyalty and revenue. For experimentation this may translate to sessions per user and revenue per user. The search engine Bing has defined their experimentation OEC to be sessions per user.

However, they have found it hard to see statistically significant improvement in this metric for experiments.

The **Primary Metric** for an experiment should be one that is supportive of the long-term goal and OEC of the site but may be more specific. For example, an experiment for a travel booking site that is attempting to increase the number of reviews of properties may have as its PM number of new reviews entered per user while at the same time not hurting number of bookings or revenue per user. An experiment that is conducted on emails sent to subscribed customers may have its PM to be percent of users that open the email (for experiments on the subject line) or clickthrough rate for number of users that open the email if improving the email message is the objective of the experiment.

Key business metrics are metrics that the business is keenly interested in seeing improved. There could be many key metrics, but it’s best if the list is not too long. Any experimental treatment that degrades one or more key metrics needs strong justification prior to implementation. For an ecommerce site key metrics may include number of days a user visits the site, sessions per user (or per day), number of categories visited per session (or per day), number of items added to cart, percent of users making a purchase (i.e. conversion rate), number of items purchased per user, revenue per user.

Diagnostic metrics are those that help determine if the experiment is harming a non-business metric or if the experiment is not working as intended. The first type of diagnostic metric is typically regarding the health of the website such as performance metrics (e.g. page load time) or number of errors users receive. The second type of diagnostic metric is to assess quality or trustworthiness of the experiment that may be due to set-up, instrumentation or interactions that may affect experimental results.

The most important of the experiment quality metrics is known as the Sample Ratio Mismatch (SRM). This is a simple statistical test that determines if the relative number of users seen in the variants is close enough to what was planned. If you have only one quality metric, it should be the SRM.

One metric that many websites see as key to success is number of visitors to the site. However, this is not a feasible metric to use in an experiment on the website to test a change to status quo. If there is a large number of new users coming to the site due to the Treatment on the website, each new user will be randomly assigned to Treatment or Control hence we will not see a difference between the two variants. However, if the test is run on another channel, such as an email test or another website, we can determine if more users came to the website due to the Treatment in a test on the other channel.

Sometimes we cannot measure the quantity we would like to measure for our Primary Metric. In an early experiment we ran at Microsoft we were testing a total redesign of the Office homepage.

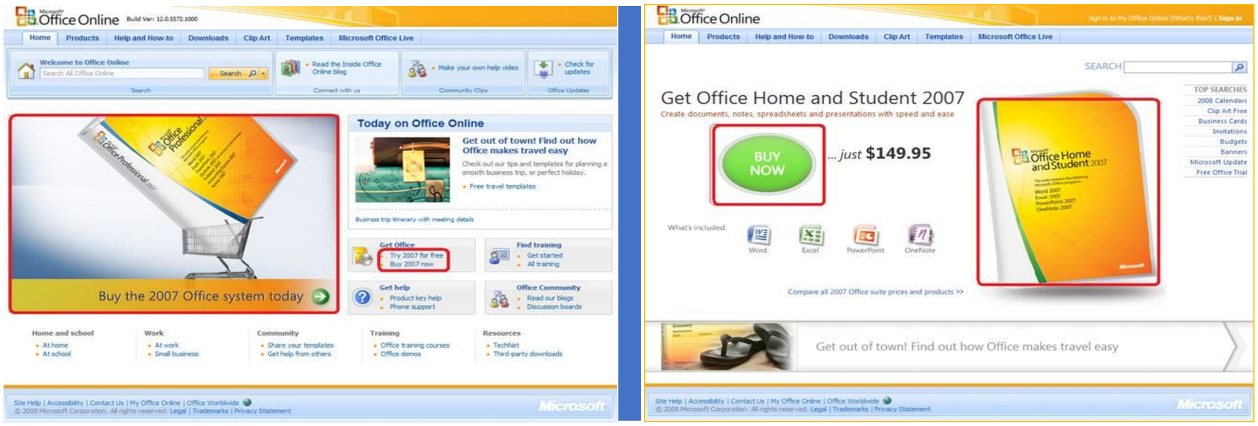

Office Online
Home
Products
Help and How-to
Downloads
Clip Art
Templates
Microsoft Office Live
Welcome to Office Online
Search
Today on Office Online
Get out of town! Find out how Office makes travel easy
Get Office
Try 2007 for free
Buy 2007 now
Find training
Get help
Office Community
Buy the 2007 Office system today
Microsoft
Get Office Home and Student 2007
Create documents, notes, spreadsheets and presentations with speed and ease
BUY NOW
... just $149.95
What's included:
Word
Excel
PowerPoint
OneNote
Office Home and Student 2007
TOP SEARCHES
Compare all 2007 Office suite prices and products >>
Get out of town! Find out how Office makes travel easy
Microsoft

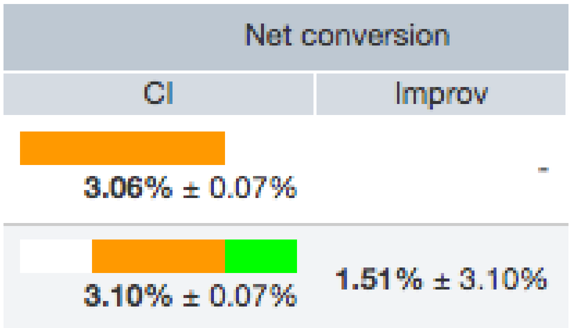
Figure 1 Control (current page) on the left, Treatment (redesigned page) on the right.

There are many differences between these two variants and the key determinant of success of the Treatment (i.e. PM) was number of Office downloads. However, once a user clicked within the red boxes on either variant they were sent to the Download Center in Microsoft. We were not instrumented to get data by experimental user from the Download Center, so we used a surrogate: percent of users to click on an area of the page to start the download process. (Note the red boxes were not on the actual webpage but are given here to show the clickable areas of the pages that lead to a potential download). In a stable situation, there is a reliable connection along the purchase/download funnel between number of clicks taking one to the Download Center and the number of downloads. For this experiment the Treatment had 64% *fewer* clicks to start the download funnel. Why did this happen? That is not a question an experiment can usually answer. We can reliably tell whether the Treatment is better, worse or approximately the same as Control but the experiment cannot tell us *why* (but it may give us clues). There are several differences that could lead to a difference in number of clicks but in my opinion, the main reason is the price is shown for the Treatment but not for the Control. When I ran experiments on Amazon where we showed the price for some electronic items versus a message that read “Add to cart to see the price” we got many more cart adds for that variant but not necessarily more sales. I believe the same phenomenon is happening here. More users are starting down the download funnel in the Control but many (perhaps most) drop out once they see the price. I believe the Treatment users who start down the purchase funnel are actually better qualified than those in the Control. Later, when we were integrated with the Download Center, we actually saw the Treatment was better. The moral of the story (as far as metrics is concerned) is to be cautious when using surrogates (or proxies). The change made by the Treatment may also change the relationship between the surrogate and the true metric of interest.

Presentation of Results

When presenting the results of an experiment there are a number of things to keep in mind.

1. There are potentially hundreds of metrics that are being compared, so succinct and visual presentation of results is important as well as effective organization.
2. The OEC, Primary Metric and other Key Metrics should be placed in a prominent position so the experimenter does not have to sort through all the metrics to find them.
3. Use color coding of results to indicate if the mean of the Treatment (or other statistic being compared) is better than, worse than or same as that for the Control.
4. Numerical results.
    a. Some basic information is needed to give the results context: number of users in each variant, Primary metric and OEC. The metrics should be grouped, e.g. by Main or Primary metrics, Diagnostic metrics, Other metrics. Within each of these there may be groupings or breakdowns into portions of the website. For example, Amazon may have Home Page metrics, metrics for each category (Grocery, Apparel, Sports,…), etc.
    b. Give the actual means for Treatment and Control as well as percent change.
    c. Also give a p-value (probability of getting a difference this large or larger if the Treatment made no difference). If the p-value is less than 0.05 it is statistically significant. If it is less than 0.01 it is very significant and should not be ignored.
    d. Finally give a confidence interval for the percent difference. (Be sure to use Fieller’s theorem to calculate this confidence interval, see later section).
5. Graphical results. I believe it is helpful to give a graphical depiction of the confidence interval for percent change. I have seen several versions of this. If the organization is using a decision point (alpha) of 5% for the hypothesis test comparing Treatment and Control, then the confidence interval(s) should be 95% confidence intervals. Several alternatives:
    a. Show the confidence interval for the mean of the Control and the confidence interval for the mean of the Treatment. The problem with this graph is correct interpretation. If there is no overlap of the two intervals the result is unambiguous – there is a statistically significant difference between the means of the two variants. However, if there is overlap the two means could be statistically significantly different if the amount of overlap is sufficiently small. Given the large number of individuals in an organization that may be interpreting these results, I believe this is not clear enough.



Net conversion	
CI	Improv
3.06% ± 0.07%	-
3.10% ± 0.07%	1.51% ± 3.10%

Figure 2 Example of graphic with CIs for Control and Treatment

b. Show the confidence interval for the difference in the two means and compare to zero. If zero is outside this confidence interval the two means are statistically significantly different, otherwise, they are not significantly different.
c. Show the confidence interval for the ***percent*** difference in the two means and compare to zero. If zero is outside this confidence interval the two means are statistically significantly different, otherwise, they are not significantly different. This is the graph I prefer since most people can better relate to percent change than actual change for most metrics. (e.g. if average clicks per user increases by 0.1, that may not seem like much of a change until you realize this is a 5% increase.)

Randomization and Analysis Units

We discussed the randomization unit in Part 1, but the choice has a profound effect on which metrics can be calculated and how analysis is carried out for different metrics when the randomization unit is different from the analysis unit. The analysis unit is defined to be the basis for a specific metric. For example, one may want to calculate clicks per user or clicks per session or clicks per pageview. Each of these metrics has a different analysis unit, namely per user, per session or per pageview. First, a review of the randomization unit, then how that affects the analysis for analysis units.

Most sites use the user or visitor ID as the randomization unit. Other sites may default to the randomization unit being a session, pageview or search (for a search engine). While there may be some positive aspects to the alternative randomization units (such as increased power for some metrics, see Part 3), there are several negatives to choosing a randomization unit other than the user.

1. If users are re-randomized into a different user experience frequently, it may be disconcerting, depending on the nature of the change. If the change is visible or obvious it could hurt the user experience for all users in the experiment.
2. For the sites I have worked with, the most important metrics are the “per user” metrics which tell us how the user performed or reacted over the duration of the experiment. (e.g. revenue per user, sessions per user, downloads per user, etc.)
3. If we randomize by user we can still calculate metrics such as revenue per session, revenue per pageview, etc. but if we randomize more frequently we cannot calculate the “per user” metrics.

The negatives of choosing user as the randomization unit are

1. The randomization is actually by the user ID, typically stored in a cookie. If the user’s cookie is deleted or if the user returns to the site using a different browser or device they will appear to be a different user and will be re-randomized into the experiment.

2. More than one person may be using the computer, in which case the results for that User ID would be for all those users.
3. If a user does not allow cookies to be set they would not be in the experiment.

If the site or experiment requires the user to be authenticated (i.e. signed in) then you could use a permanent User ID associated with that account. Obviously, this is the preferred method for randomization but it only applies to a minority of sites/experiments.

It is my opinion that the pros for randomization by user outweigh the cons for most experiments.

The choice of randomization unit affects the analysis for metrics in the following way. First, is it common to assume the Stable Unit Treatment Value Assumption (SUTVA) (Imbens and Rubin 2015), which states that randomization units (e.g., users) do not interfere with one another. If this is true, then we can treat observations for a user as uncorrelated with other users. This makes the analysis of metrics which have an analysis unit the same as the randomization unit straightforward. However, when the analysis and randomization units are not the same one must use another approach to calculate standard deviations. Two popular methods are the delta method (Deng, Knoblich and Lu, 2018) and bootstrapping (Bradley and Tibshirani 1993).

Analysis of Ratios

Comparing the means for Treatment and Control along with a confidence interval for the difference is straightforward. However, when showing the amount of change to the experimenter, the percent change usually is more meaningful. Saying the Treatment is 2.5% greater than the Control has some value but not as much as including a 95% confidence interval for the percent change. Then the experimenter would have much better information to make a decision. For example, the three confidence intervals in Figure 4 all have a 2% increase of Treatment over Control.

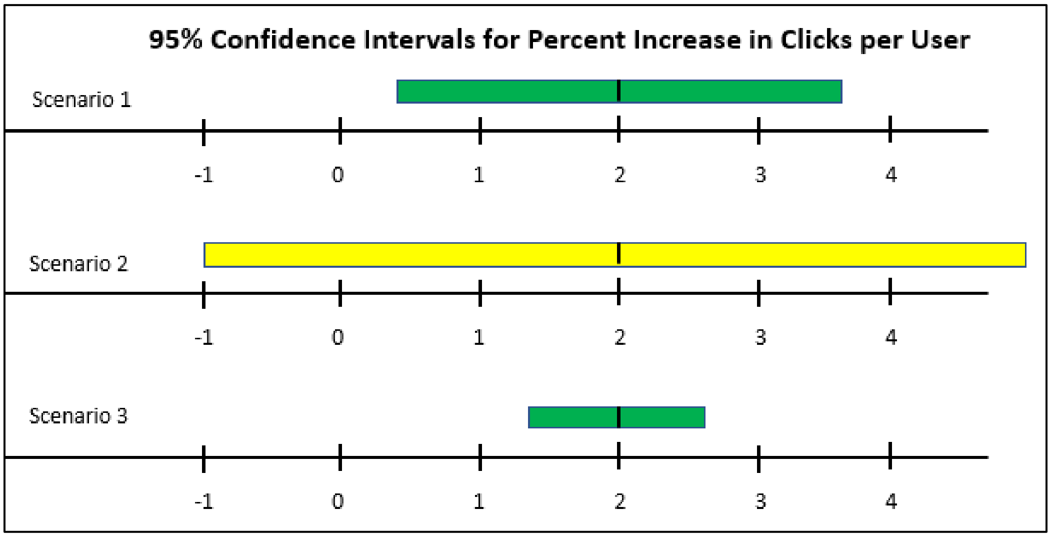


Figure 3 Three Potential Confidence Intervals for Percent Increase in Clicks per User

In Scenario 1, the experimenter can see that the Treatment is statistically significantly better than Control (since zero is not in the interval), but only slightly. In Scenario 2, the 2% increase is not close to being statistically significant, whereas in Scenario 3 the difference is highly significant and the experimenter can say with a high degree of certainty the percent increase is greater than 1%. (One would not see these three scenarios in the same experiment at the same time, but could see a progression from Scenario 2 to 1 to 3 as the sample size increases while the experiment runs longer.)

The complexity comes in when actually calculating the confidence intervals for percent change. One must use Fiellers theorem (Fieller 1954) to get it right. Not all experiment UIs will show a graph for the confidence intervals, but should at least state the upper and lower bounds.

Initial experiment diagnostic checks

Regardless how careful the experimenter is or how well the experimentation platform is constructed, experiments will have problems. When I was running experiments at Amazon (many years ago!) I had an experiment that had a puzzling result. This was a high profile experiment with many management eyes on the results. Everyone thought the Treatment should be better than the Control but the results were not turning out that way. In fact, it was highly significant in the negative direction. I began my investigative work to find out. One issue I uncovered was a slight difference in the number of users in each variant than we expected. The allocation was planned at 50% in each variant but we had 50.23% in the Control and 49.77% in the Treatment. Some wanted to dismiss as “not material” and even if was statistically significant we were randomly allocating users to each variant, so why would this cause a problem? In fact, a small imbalance in number of users can have an out-sized impact on the test results. In this case, the difference from expected was highly statistically significant (p-value<0.0004). I went on to find that the Treatment was losing some users causing a significant bias in the metrics. We later made this calculation a standard diagnostic test and most experimentation platforms use this diagnostic today. I wish I could say “If you see this diagnostic signal there is a problem, and THIS is the cause.” In fact, this is more like a child having a high fever. You know there is a problem, but there are many possible causes. You will have to do some investigative work to find where the imbalance is coming from. An SRM can be the result of many types of problems that may be due to instrumentation, triggering, lost telemetry, redirects, performance issues, and other causes. Microsoft estimates 6% of their experiments exhibit an SRM and LinkedIn reports 10%. A good treatment of how to investigate this signal is given in the article (Fabijan, Gupchup, et al. 2019). The book, *Trustworthy Online Controlled Experiments* (Kohavi, Tang and Xu 2020, Chapter 3) give a thorough treatment of this and other diagnostic tests.

When you set up your diagnostic metrics, the Sample Ratio Mismatch (SRM), described above should be the first to deploy. It is easy to do and exposes a surprising number of underlying problems that could be biasing your experiments to give you incorrect results. Anytime you see this signal you must take it seriously. If it fires, there is a problem. However, since this is calculated for every experiment you don't want to set the threshold too low. My recommendation is to sound an alarm if the p-value is less than 0.01 and raise a serious red flag if the p-value is less than 0.001. You don't want experimenters to deploy their treatment if you are getting a very small p-value for this diagnostic!

Another set of quality metrics looks for a large difference in Treatment effect over time or among subgroups of users. The next figure (also shown in Part 1 of this series) illustrates a change over time that you need to be alerted to.

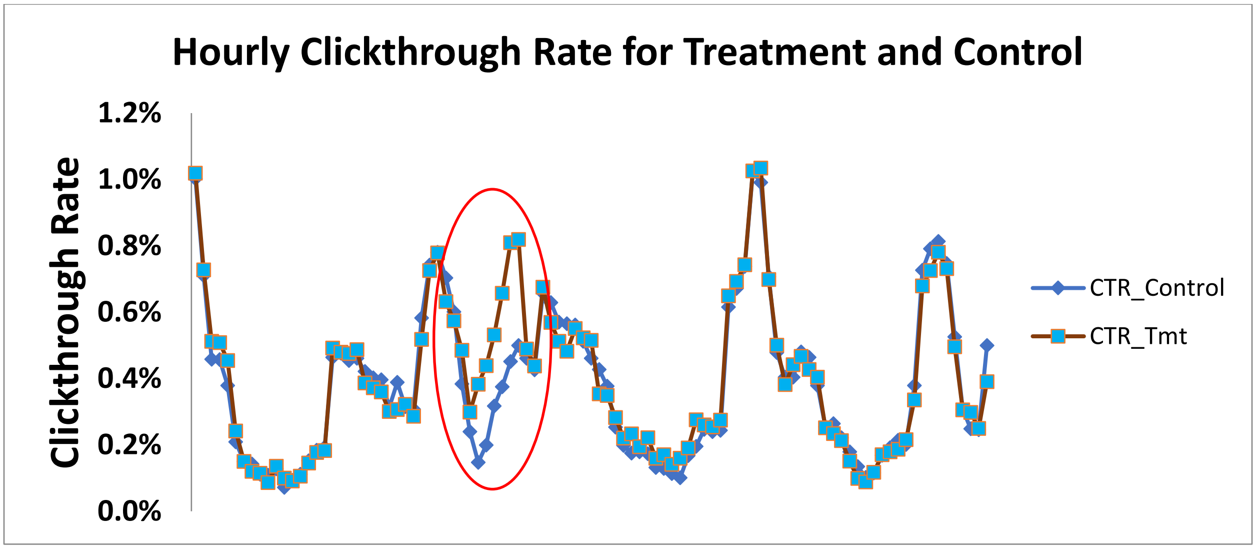


Figure 4 Comparison of Clickthrough rate for Treatment and Control over 4 day period

In this experiment the content of the site should have been the same for both variants but Treatment had a different headline from the Control for a 7 hour period, biasing the results.

A related diagnostic that should be deployed early is to compare the treatment effect for different segments of your user population. For example, we found a number of problems with browser data collection when we segmented results for the different browsers. Other segmentations of interest may be by device type, language, country, etc. A good overview of heterogenous treatment effects is available at EGAP (2018). Identifying interesting segments, or searching for interactions, can be done using machine learning and statistical techniques, such as Decision Trees (Athey and Imbens 2016) and Random Forests (Wager and Athey 2018).

Another good rule (although not strictly a test) is to distrust any result that is very surprising, whether it is surprisingly bad or surprisingly good. You should not rejoice when getting a huge increase in Treatment without doing due diligence to make sure it is a trustworthy result.

A/A tests

First, a definition. An A/A test is like an A/B experiment except there is no change made for the second variant. Someone who is not familiar with A/B testing may expect there should be no statistically significant results when you run an A/A test. In fact, you should expect about 5% of the metrics to be statistically significant. (This is assuming you use the 5% Type I error rate. Some sites use a lower rate and others advocate for a larger Type I error, but I think the 5% error rate is good if you use it appropriately.)

What is the purpose of running an A/A test? There are actually several objectives. (See Kohavi (2020) for more in-depth discussion.)

1. As a diagnostic for your experimentation system.
    a. When first setting up your experimentation system an A/A test should be run on all visitors coming to the site. The number of observations (clicks, pageviews, etc.) found in the experiment should be compared to the number of observations counted by the existing system of record. It is unlikely you will get the same number, but they should be close. If not, investigate.
    b. When you have many metrics calculated from an A/A test you expect the distribution of p-values to be uniformly distributed between 0 and 1. One example of when this would not be true is if the calculation for standard deviation for some metrics is incorrect. Another is if an incorrect assumption is made for the hypothesis tests (e.g. if two-sample t-tests are used to compare the means and the means are not close enough to being normally distributed due to skewness or outliers.)
    c. Detect some biases that may exist between variants. One example is if the treatment group of users is the same as the treatment group in a previous experiment. If there is a “carryover effect” (positive or negative) it could bias the next experiment. If your system is set up this way you should always run A/A tests between experiments.
2. Estimate standard deviation
    a. With an A/A test that is run for four weeks you can calculate the standard deviation for all your metrics for one to four weeks to see how the standard deviation changes as more users are added to the experiment. For some metrics, as you run an experiment longer and more users are added, you get an expected reduction in standard deviation of the mean. However, for other metrics, the standard deviation may not reduce much, if at all. (See *Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained* (2012) for discussion of the metric sessions per user.)
    b. The results from the change in standard deviation over time for your primary metrics in an A/A test can be used to determine how long you should run the

experiment to achieve a specified power. Briefly, the power of an experiment (for a specific metric) is the probability of detecting a change of a certain size due to the Treatment. (A more in-depth discussion of power will be given in Part 3.)

Topics for Part 3:

- What is Power and how do I get more?
- How long should I plan on running this experiment?
- When and how to ramp up an experiment
- Ending an experiment early
- Platform options
- Ethical considerations

I hope you have much success in optimizing your website, email campaigns, mobile apps, etc. using A/B testing. If you want to discuss how to get started or how to improve the testing you are currently doing, please contact me. I'll be glad to help.

References

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *PNAS: Proceedings of the National Academy of Sciences.* 7353-7360. doi:https://doi.org/10.1073/pnas.1510489113.

Bradley, Efron, and Roberg J Tibshirani. 1993. *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Deng, Shaojie, Ulf Knoblich, and Jiannan Lu. n.d. "Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas." *24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* Association of Computing Machineray.

EGAP. 2018. "10 Things to Know About Heterogeneous Treatment Effects." *EGAP: Evidence in Government and Politics.* egap.org/methods-guides/10-things-heterogeneous-treatment-effects.

Fabijan, Aleksander, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer, and Pavel Dmitriev. 2019. "Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners." *KDD '19: The 25th SIGKDD International Conference on Knowledge Discovery and Data Mining.* Anchorage, Alaska, USA: ACM.

Fieller, E. C. 1954. "Some problems in interval estimation." *Journal of the Royal Statistical Society, Series B* 16 (2): 175-185. doi: JSTOR 2984043.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Kohavi, Ron, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* Cambridge University Press. https://experimentguide.com/.

—. 2020. *Trustworthy Online Controlled Experiments.* Cambridge: Cambridge University Press.

Kohavi, Ron, Roger Longbotham, Sommerfield Dan, and Randal M. Henne. 2009. "Controlled experiments on the web: survey and practical guide." *Data Mining and Knowledge Discovery* 18: 140-181. http://bit.ly/expSurvey.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal fo the American Statistical Association* 13 (523): 1228-1242. doi:https://doi.org/10.1080/01621459.2017.1319839.