
# What I learned from Udacity's course on A/B testing, by Google

Nikhil Sawal Nov 5, 2018 · 16 min read

In the phase of evolving websites and mobile applications, A/B testing continues to be one of the most commonly used techniques and a to-go choice for managers and decision makers for answering business questions. A/B testing helps you quantify, the way users respond to new product or a new feature by comparing it with the original version and seeing which version performs better.

Through this article, I wish to convey my understanding about Udacity's course on A/B testing, by Google. I would recommend this course to anyone considering a career in data science. But before you take this course, be sure to have your concepts on

course or if have the luxury of time, you can do Intro to Inferential statistics, another very good course on Udacity. Without any further due, let's get started. Following is how the entire article is structured.



## 1. First things first! What is A/B test?

commonly known, Design of engineering experiments. The experimental units (incoming traffic to a web page or the users) are divided into two groups (A and B), where one group is exposed to the original version and the other group is exposed to the new version. The end goal is to evaluate success metrics and decide whether to launch the new feature or not.

## 2. Why are A/B tests important?

A/B test are important for 2 reasons:

1. They are a great way to cater short term business questions.
2. A/B tests help draw causal conclusions

## 3. Why A/B tests? Why not controlled experiments?

A/B tests are like controlled experiments or split tests, which include hypothesis, control and treatment groups (or original vs. the variation that you want to try) and statistically calculated results. The only difference is that, when an experiment is conducted in a controlled environment, it is referred to as a controlled experiment whereas, A/B tests are experiments conducted on the internet. The reason why A/B tests are not commonly referred to as controlled experiments is that, there is a lot of variability associated, when it comes to internet traffic, which could be mitigated to a certain extent, but would be difficult to operate in a completely controlled environment.

## 4. Are A/B tests just limited for checking two variations, as the name implies?

NO! A/B tests are not limited to just two variations, you can try out n number of variations. But as n increases, the tests demand more traffic per group. Other types of online experiments involve:

- Multivariate Tests
- Bandit Algorithm

I will briefly touch upon each of these.

of each of the 4 pages) are the attribute and the 4 different colors represent the variations that we are testing for. We can see that the yellow variation is doing better than the rest of the colors.

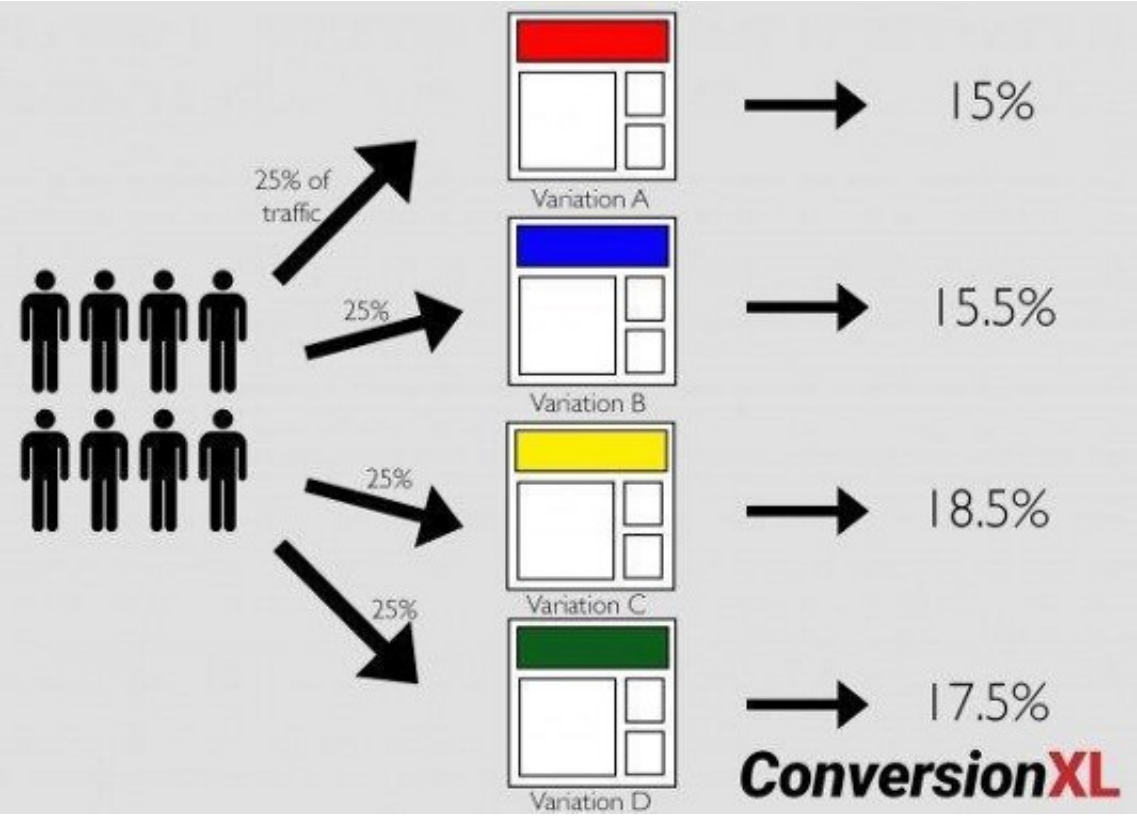


Image source

**Multivariate tests:** Tests multiple attributes each with multiple variations. *MVTs are used to determine which set of attribute variations or attribute combinations, produce the best results.* In the following figure Item 1, Item 2 and Item 3 represent the three attributes, each of which has three variations represented by blue, pink and gray blocks. Use multivariate tests to polish the layout of the page.

Image source

**Bandit algorithm:** Bandit algorithms are A/B/n tests that update in real time based on the performance of each variation.

## 5. What are the phases of an A/B test?

There are 2 important phases in any A/B test:

1. Exploration: This is the phase when you test the change you want to launch, on a small but representative sample of the target population, before deploying the change, whilst giving careful considerations to avoid the risk of a type I error (false positive) and type II error (false negative). Quite a big statement! Let's break it down into small chunks and try to understand.

- "small but representative sample of the target population": Having a representative sample is important for your conclusions to be deemed valid. Before you launch your changes to the entire target audience, you would want to try out your variations on a *small but representative* sample of the population, so that you can observe the response of the smaller audience and if the outcome is good, extrapolate your results to the entire target audience.

- "avoid the risk of type I error (false positive) and type II error (false negative)": type I error (false positive) refers to, concluding that a change or intervention has an effect, when in fact it DOES NOT. Type II error (false negative) is failing to launch a change that is significant. The following figure shows an example of a false positive.

Image source

2. Exploitation: Launching the changes. Pretty self-explanatory right?

## 6. What to A/B test and what NOT?

A/B testing is good for answering "this or that?" kind of questions. Some examples are:

- Does increase in the page load time, impact revenue? Here you might have made some changes in the user interface (UI), which might have increased latency. So, you would want to A/B test the original UI vs. the new UI and see if there is a difference in the revenue.
- Another example can be, testing if changing the color of a button, improves the click rate? This is used when you want to test the usability of the button, to see if the button is noticeable in its current shade and size. A similar experiment was run by Google when they were not able to decide between two blues, so they ended up testing 41 different shades of blues, between the two blues.

On the other hand, A/B testing may not be a good choice to check things like, the completeness of the website, new experiences (like change aversion vs. novelty effect)

- Completeness of website: A/ B testing cannot answer questions like, do we have a missing product on our website? It can answer if product A should be above product B, but it can't tell if there is a product C that is missing from the website and that we should include it or not.

- New experience: Some users might not be happy with new changes (this is change aversion) while other users might be excited about trying out the changes (this is called the novelty effect). So, it will be difficult to come up with a baseline, in that you won't know what percentage of your incoming traffic belongs to change aversion and what percentage belongs to novelty effect. That makes it harder to come up with a comparable control/treatment split. The other issue is the timeline, the time needed for users to adopt to the changes, so that you can make robust conclusion.

- Long-term changes: These could be hard to measure as well. Let's say you want to check the effectiveness of a referral scheme for rented apartments. Now people don't rent apartments that often, which makes it difficult to gather enough data within a short period to make valid conclusions. Because the longer you run your experiment, the greater is the chance that the change you are observing is caused by a factor, that came up in recent times, but was not included at the start of the experiment. For example, assume you run the experiment of rented apartments from May to August. Now the month of August, typically marks the beginning of a semester, which means that you might observe a sudden surge in the number of apartments rented. So, if you fail to consider this factor, you might just have a false positive.

## 7. Steps for a well-structured A/B testing

### a. Research

Before you think of doing any experimentation or optimization, just take a moment and think

- Do you have a clear idea of the business objective?
- Do you have a well-defined goal?
- Do you know what KPI (Key Performance Indicator) are you going to use?
- Do you have a target metric?

Once you are clear about these things, you can move on to doing some sanity checks like:

- Are there any bugs in the website? Bugs can be potential conversion killers.
- Are there any discrepancies across different devices and browsers? These discrepancies could be latency issues across different devices and browsers. Conducting tests to check consistency across different platforms, could potentially obliterate the risk of having a confounding or lurking variable ruining the experiment.

### b. Choosing and characterizing a metric

The choice of metric depends on the purpose you want the metric to serve. There are two classes of metrics:

- Metrics used for evaluation: These metrics include the detailed metrics that help track in-depth information like user experience with products and then there are high-level business metrics that are used to track business objectives.
- Metrics used for sanity checks: These metrics are used to check if the control and experiment groups are comparable. So, you may check if you have the same number of users in both groups. Are the groups comparable in terms of distributions and so on?

Steps to come up with a metric

***Step 1: Come up with a high-level concept for the metric***

This is usually a definition that everyone agrees with. These metrics can be as simple as click-through-rate, click-through-probability, number of active users

So, we saw in step 1 that — number of active users — is a high-level metrics. A fully realized version of it would be to define what active means. Is it a 1-minute active? Or 1 hour active? Or 1 week active? So, one thing to keep in mind when defining a fully realized metric is to decide on the time-stamp. The other thing you might want to consider, is the consistency across different browsers. Each browser might have a different way of interacting with your website and hence might be capturing data differently. Also, the latency might be different across different browsers and this might cause discrepancies.

Now the effectiveness of these metrics, bank heavily on the data you use for computing them and hence, evaluating different techniques and sources of gathering data, deserve equal attention.

***So, what data to use?***

You would obviously be using data generated from your website like clicks, time spent on a page, page views, number of accounts created and so on to run an A/B test. But in addition to this data you can use external data generated by companies that focus on gathering data to answer interesting questions through surveys or you could hire companies to generate your own in-depth data using techniques like user experience research and focus groups. Data gathered from these techniques can serve as a good source to validate your metrics. This can be achieved by plotting trends and seasonal variability for internal data VS. external data and checking if they line up. Second, you can also use this data to generate ideas about using simple-to-measure metrics, as proxies for harder-to-measure metrics. An example of a harder-to-measure metric could be, measuring user satisfaction. In a study conducted, user reported satisfaction was compared to time spent on the website to establish a general correlation between the two. And through this correlation — duration or time spent on a website — which is easy to measure, was converted to a metric that was used as a proxy to measure user reported satisfaction.

***Step 3: Summarize measurements into a single metric***

need to consider the following two characteristics of our final summary metrics:

1. ***Sensitivity and robustness:*** We want our metric to be sensitive enough to capture the changes we care about, but at the same time the metric should be robust enough to not capture the changes, that we don't care about. Sensitivity and robustness are important because they save us from type-I error (launching a change we don't care about) and type-II error (failing to launch a change we care about). To measure sensitivity and robustness we can:

- Run experiments: If we have access to experimental data, then we can run experiments and see if the metric moves in conjunction to the changes. This would be a good test for sensitivity.
- A vs. A test (or A/A test): We can also use the experimental data to do an A vs. A test. In an A/A test, both control and treatment groups are exposed to the exact same change. If our metric moves across the two groups, then this would be a good indication, that the metric is not robust enough.
- Retrospective analysis: Now if we don't have access to experimental data, then we can pull out some data that we archived, which was used to run a similar experiment in the past and see, if the metric that we are interested in, responds in a way we intend it to.

***2. Distribution of the metric:*** Distribution is used to prioritize one particular metric over the others. Let's say, we have a nice well-behaved distribution like the normal distribution, then mean or median would be a good choice. If the distribution is a skewed one, then percentiles would make a good choice.

## c. Choice of experimental unit and population

The next important topic to consider is the experimental unit (Unit of diversion) and the population you wish target. Both are increasingly important when it comes to sizing the experiment and deciding how to assign events to either the experiment or control group.

1. ***Experimental unit:*** This is the unit of diversion used to define which user or which event is assigned to the control and experiment group. The unit of diversion can

important considerations:

- User consistency: If we are dealing with a user-visible change, we would want our user to have consistent experience throughout. So, a user_ID or a cookie would be good choice, where as if the change that we are trying to implement, is not user visible, then an event-based diversion like a page view would make more sense. This is important, because if we use page views as a unit of diversion for a user visible change, then every time the user reloads the page, they might be assigned to a new group, i.e. if the user was initially in the experiment group, now they may end up in the control group.

- Ethical considerations: Since actual people are involved in the experiment as experimental units, it's very important to give careful considerations to the ethics of the experiment. Some ethical considerations are risk, benefit and privacy. If the risk exceeds the threshold for the minimal risk, i.e. it encompasses physical, psychological, emotional, social or economic concerns, then getting an informed consent becomes crucial. If the users would benefit post completion of the study, then stating the benefits is important. If the internal processes for collection of new data are well in place, then privacy won't be a huge issue, but if not, additional safety measure would be needed.

- Variability of metric: The choice of unit of diversion can greatly impact the variability of a metric. The variability of the metric is much higher if the unit of diversion is broader as compared to the unit of analysis. Unit of analysis is basically the denominator of the metric. So, for click-through-rate, which is defined as #clicks / #page views, #page views becomes the unit of analysis, (where '#click' is read as 'number of clicks'). So, if we use user_ID as our unit of diversion and click-through-rate as our unit of analysis, the variability of our metric, click-through-rate would be much higher, since one user_ID can correspond to multiple page views.

2. ***Population:*** Choice of population will greatly impact the success of the new change or the new feature, you are trying to implement. When you are planning to launch a change, you would want to launch the change to the most relevant audience i.e. user

- Avoid unwanted media coverage: This is particularly helpful, when you are not sure if you are going to launch the change. So, you might want to limit the number of users exposed.
- Second reason is that if you are planning to release the change internationally, you might want to be extra sure that the language used is right.
- Next, if you know that the change is going to impact users belonging to a certain demographic, then releasing the change to the entire audience might delude the effect of the experiment.

Now the main question!

***How does choice of metric, unit of diversion and population impact the size of the experiment?***

- We saw in the previous section that if the unit of analysis is the same as unit of diversion, the variability of the metric is reduced. Now since the variability is reduced, the number of page views needed for the same minimum detectable effect also goes down.
- Targeting population correctly also reduces the sample size required for the experiment, because we are not considering audience that are irrelevant as it might delude us.

### d. Duration of the experiment

The duration of the experiment is related to the fraction of the traffic you wish to send through the experiment. If you need a total of 20K users for your experiment (10K users in each control and experiment group), sending 500 users per group per day, would require you to run the experiment for 10 days. Now if you reduced the number to 250 user per group per day, the duration for the experiment would go up to 20 days.

The next important thing to keep in mind when it comes to duration is that, *Statistical significance does not imply practical significance!* Just because your test showed statistically significant results doesn't mean you should launch the change. Statistical

That being said, lets formally state some stopping rules:

> *Stopping Rule:*
>
> *Decide on the duration (typically 2–4 business cycles). The reason we consider full business cycles is to avoid skewed results and gain a representative sample, because it will include every external factor: every day of the week, different traffic sources, seasonality and any other external events, twice.*
>
> *Decide on a minimum sample size per group per day (like 400 users', per treatment and control group)*

One important phenomenon that might come up is regression to the mean. As per this phenomenon, you might observe a clear winning variation in the initial phase, but as the test progresses, there are no differences in the conversion rates. The figure below shows the conversions, regressing to the mean after 4 weeks. So, if you observe diminishing differences between different variations during the course of the test, then this may be one indication why you should run tests longer. It also saves us from launching a variation which is no different than our current version.



Image source

## e. Analyzing the results

The first step in analyzing any experiment is to check if the control and experiment groups are comparable, basically do the sanity check using the invariant metrics that we discussed section 2 (Choosing and characterizing a metric). If the sanity check fails, there is no point in proceeding with the rest of the experiment. Now the two groups may not have the exact same number of experimental units, but they should be roughly comparable.

***Analyze results***

If you have a single evaluation metric, then you can directly construct a confidence interval for the difference that you observed. If the confidence interval does not include 0, then the difference observed between the control and treatment is statistically significant but as we discussed early, statistical significance does not imply practical significance, if the minimum detectable difference is lower than confidence interval, then we can say that we have observed a significant difference that we care about.

We can compare our results to those obtained from non-parametric tests like sign test and see if they align with the results we observed for our hypothesis.

Now what if we don't observe statistical significance?

Don't move on too quickly. A/B testing is an iterative process. Try a few more iterations.

Second, look for Simpson's paradox. Just because your overall test didn't show statistical significance, doesn't mean your intervention (or change) isn't worth launching. The change might have improved the conversion rates for a specific segment of audience. In this case, you can make the new version available to that segment, instead of launching it to a wider audience, though you might want to make sure that you can justify, why the change improved experience for that segment. So, lookout for segmentation.

Some of the segments which you might consider are as follows:

- New users vs. Experienced users
- Mobile/Tablet vs. Desktop

- Demographics (age group, gender, city, state, country)
- Users within vs. Users beyond the sign-in/sign-out boundary
- Traffic landing directly from the page vs. Traffic coming via internal link

Now, if you have multiple evaluation metrics, chances are that, there exist some correlation between the metric, but you can always assume independence and get a conservative estimate, so you can use something like the Bonferroni correction

## 8. Conclusions

The final step is to draw conclusion. To make conclusion, you basically need to have answered three key questions throughout the course of the experiment:

- Do we have statistically and practically significant results?
- Do we understand the change well enough?
- Is the change worth launching?

## 9. Key takeaways

- Check, double check, triple check the set up of the experiment
- Statistical significance does not imply practical significance
- Be vary of Simpson's Paradox and Regression to the mean
- If it's your first experiment that might have a big impact, run a couple of experiments and see if you are comfortable with the outcome

## 10. Summary

- Follow a structured approach
- Research and define the business objective or the KPI that you want to improve.
- Conduct the experiment
- Analyze, learn and iterate

- Statistical significance does not equal validity (or why you get imaginary lifts)

Data Science A B Testing Statistics Metrics Experiment