EXPERIMENTAL DESIGN

Main takeaways

Considerations before designing the experiment

Perform A/B Testing only on an **established site**, when you have a stable baseline.

Perform A/B Testing only when you have solved all of the “obvious” or “known” problems, and you are **not sure what will work** or not.

Make sure the **strategy** of the company is aligned with whatever you want to improve with the test.

Make sure management supports spending the **time** and **resources** to perform the test, and **commits to adopting the winning condition**.

How many different versions should be tested?

Fewer variants

Simplicity

Less traffic/ fewer users required

Cheaper in terms of variant creation

Fewer ideas can be tested

More variants

More ideas tested: more chances of discovery

More traffic required (longer experiment)

Statistically more complex

More expensive (the variants have to be created)

Companies learn to walk (less variants), before they can run (more variants)

Focus on coming up with interesting and well-thought variants!

What kind of changes can we implement in each variant of the test ?

Generally, you want to do only one change / variant, and only across one “dimension” at a time:

- Color of the button
- Text of the button
- Size of the button
- …

Implementation, tracking & analysis

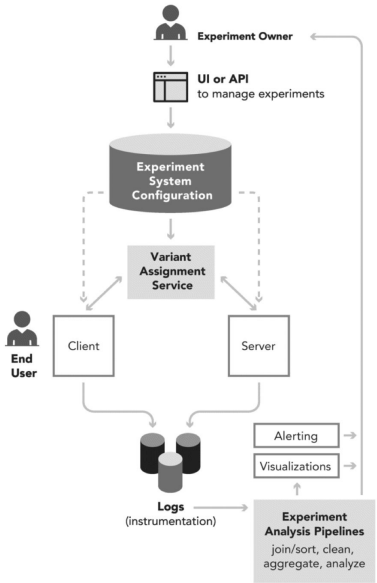
For simple front end changes:

Optimize

Optimizely

For back end changes:

- Cross platform needs
- Whenever speed is critical
- Tracking complex ad-hoc metrics
- Privacy concerns



Which metric should we choose to compare the different versions?

Our main metric or **OEC** (Overall Evaluation Criterion) must be **measurable in the short term** (the duration of an experiment) and **believed to affect long-term strategic objectives**.

The OEC should be a ratio defined as:

$$\frac{\textbf{Number of conversions (successes)}}{\textbf{Chances to convert}}$$

The strategy of the company determines the choice of OEC

A good and very common metric to use here would be **Click Through Rate** (CTR).

It's easy to measure and repeatable with many users

A bad metric to use here would be **Profit**.

It fluctuates due to many factors outside the scope of the experiment

Other metrics to ensure the validity of the OEC

Driver metrics

Indicate we are moving in the right direction towards our goal

- Happiness
- Engagement
- Retention
- etc…

Guardrail metrics

Guard against mishaps

- Latency
- Bounce rate
- Sample size
- etc…

Should Eniac experiment with other elements of the site instead (or in addition to) the “SHOP NOW” button?

Not elements on the same page! Otherwise it would be hard to distinguish between the effect of the different variants.

Many companies run hundreds of tests in parallel —it is completely fine as long as the interactions between them are controlled.

RED → Original search bar

WHITE → Original search bar

RED → new search bar

WHITE → new search bar

How can we be sure that the variant with the best performance is not having more clicks due to just chance?

We will perform a **statistical test** to compute the probability of our results being due to chance, under the assumption that all variants are equal: this is the definition of the **p-value**. If this probability is low enough, we will reject the "chance" factor (which is called the "null hypothesis").

"Low enough" is a threshold that we need to define beforehand, and it's called **significance level.** There are common thresholds (e.g. 95%), but ultimately we will have to choose one that fits our problem.

Ask yourself the question: "how bad would it be to conclude there is a real effect when there is none?". If you responded something along the lines of "it would be deathly", then you want to set a very high significance level.

How long can we expect the experiment to last?

The length of the experiment depends on the **sample size** needed: the longer the experiment runs, the more users/visits will our site receive. So, to get a given sample size, the first factor to take into account is the **traffic** received by the tested page. But the sample size itself depends on these factors:

- The **significance level**: a higher significance level (e.g. 99%), or a higher confidence that the results are not due to chance, will require a larger sample size compared to a lower one (e.g. 90%).
- The **minimum detectable effect**: if you care about detecting a small difference between the variants (e.g.variant B is 0,2% better than variant A), you will need a larger sample size. If you only care about a big difference (e.g. 25%), you will be fine with a smaller sample size. This is also called **practical significance.**
- The **number of variants**: the more variants, the bigger sample size needed.
- The **conversion rate**: if you're having many visits but nobody clicks any button (neither the A or the B), it will take longer.
- The **statistical power**: the probability that the test will correctly reject the null hypothesis.

Other considerations:

- How fast-pace is your industry?
- Is there a novelty/adoption effect that alters the results on the first days?
- Seasonality

https://www.abtasty.com/sample-size-calculator/