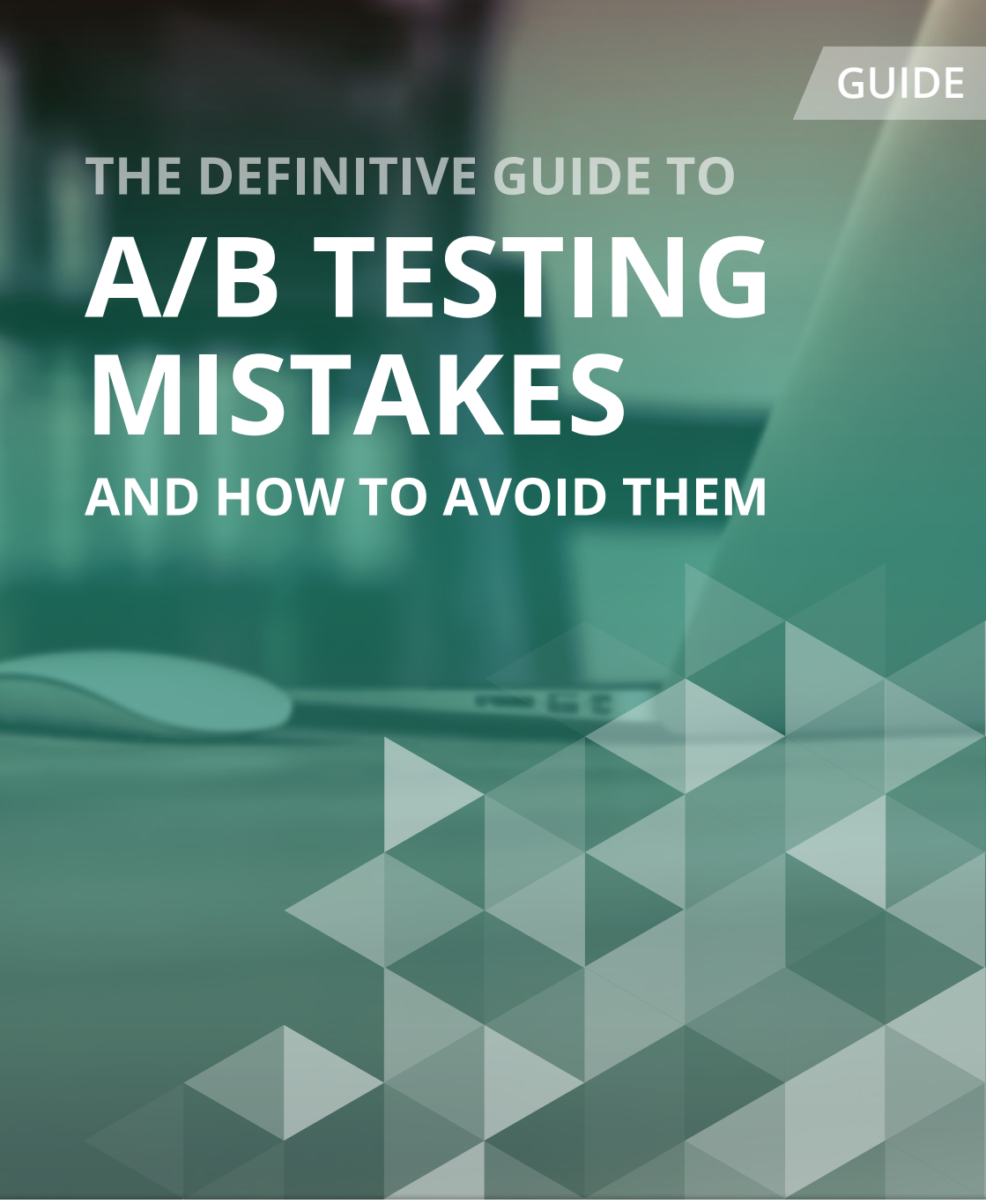
GUIDE

THE DEFINITIVE GUIDE TO A/B TESTING MISTAKES

AND HOW TO AVOID THEM



Kameleoon

TABLE OF CONTENTS

0 INTRODUCTION 5

1 7 BEGINNER MISTAKES 7

2 YOU ARE STOPPING YOUR TESTS TOO EARLY 21

3 WARNING! IS THE WORLD SABOTAGING YOUR A/B TESTS? 31

4 ARE YOU MISINTERPRETING YOUR RESULTS? 43

5 YOUR BRAIN IS YOUR WORST ENEMY 55

INTRODUCTION

Unless you've been living under a rock, you know that **A/B Testing can be extremely beneficial for all your marketing and design decisions**.

It allows you to know exactly what works, what doesn't, to test before implementing anything so you don't lose money, and more.

You're most likely already doing A/B tests.

All the power to you right? Well, not so fast. Here is the bleak truth...

A staggering number of people doing A/B tests get imaginary results. We see it among our clients, and 5 minutes on Google will tell you the same story.

How come?

Because you can screw up in so many ways it will make you dizzy, and because maths are involved.

But fear not, we got your back.

In this ebook, **we'll look into about every mistake you could make while A/B Testing.**

So grab a cup of coffee (or 3) and let's begin!

1

7 BEGINNER MISTAKES

7 BEGINNER MISTAKES

Let's start off with 7 mistakes most beginners make when A/B testing:

1. They start with complicated tests
2. They don't have a hypothesis for each test
3. They don't have a process and a roadmap
4. They don't prioritize their tests
5. They don't optimize for the right KPI's
6. They ignore small gains
7. They're not testing at ALL times

1. They start doing A/B Testing with complicated tests

For your first tests ever, start simple. Being successful at A/B Testing is all about process. So it's important that you first go through the motions.

See how theory computes with reality, what works for you and what doesn't. Where you have problems, be it in the implementation of the tests, coming up with ideas, analyzing the results, etc...

Think about how you'll scale your testing practice, or if you'll need new hires for example.

Starting with A/B Testing is a lot like starting weight training seriously. You don't start with your maximum charge and complicated exercices. It would be the best way to injure yourself badly.
You start light, and you focus 100% of your attention on the movement itself to achieve perfect form, with a series of checkpoints to avoid all the ways you get injured or develop bad habits—that'll end up hurting in the long run.

By doing that, you'll imprint it in your muscle memory so when you need to be focused on the effort itself, you won't even have to think about the movement. Your body will instinctively do it.

Then and only then, can you achieve the highest performance possible without screwing up.

Exact. Same. Thing. With. A/B Testing.

You start simple, focus all your attention on each step of the process, set up safeguards and adjust as you go so you don't have to worry about it later.

Another benefit if you start with simple tests is that you'll get quick wins.

Getting bummed out when your first tests fail (and most do, even when done by the best experts out there) is often why people give up or struggle to convince their boss/colleagues that split testing is indeed worth the time and investment.

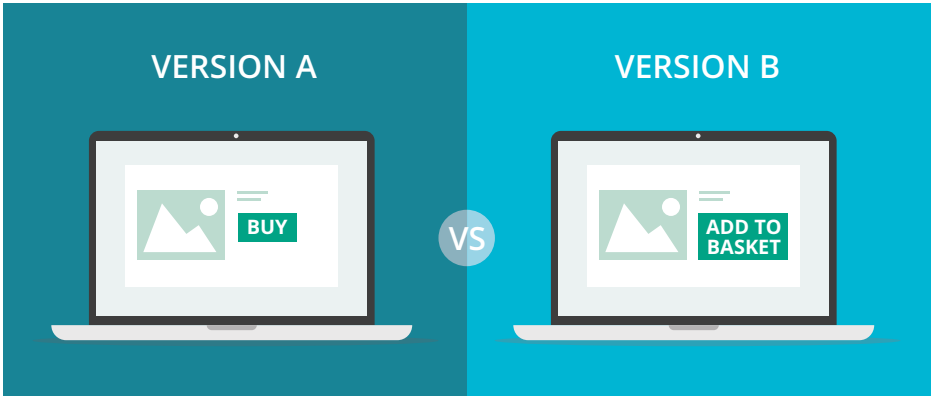
Starting with quick wins allows you to create momentum and rally people to the practice inside your team/company.

So, you got the message: doing complicated tests right off the bat could kill your efforts in the egg.

You could be overwhelmed, get fake results and get discouraged.

Here are a couple examples of things you could test to start with:

- Test copy on your offers, product pages, landing pages (Make it focused on benefits not features, be sure that what you mean is crystal clear)
- Removing distractions on key pages (Is this slider really necessary, or all these extra buttons?)



2. They don't have a hypothesis for each test

Every test idea should be based on data, articulated through an informed hypothesis with an underlying theory. If you're not doing this, you're mostly shooting in the dark without even knowing in what direction the target is.
Not having a hypothesis (or a flawed one) is one of the most common reason why A/B tests fail.

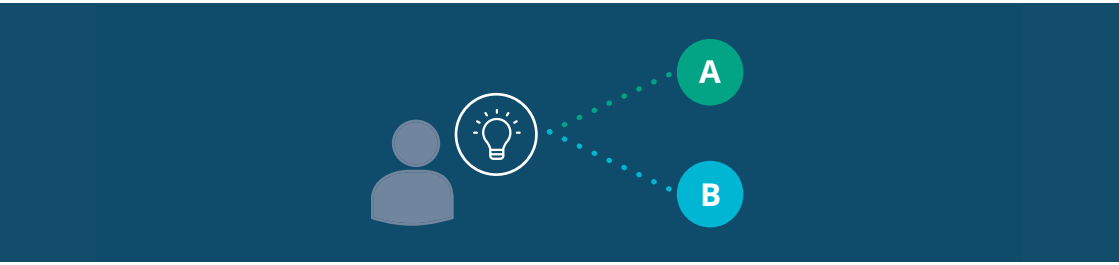
Okay, let's take our initial statement apart piece by piece to understand what this means.

Every test idea should be based on data [...]
Be it quantitative or qualitative, every test is induced by an analysis of your website data and the identification of a problem or element to improve.

Gut feelings or "I read online that..." won't do. It could work but you're A/B Testing to make decisions based on data, so let's not flip a coin for our tests ideas, shall we?

Here are several sources to build your hypothesis on:

- Analytics
- Heatmaps
- Surveys
- Interviews
- Usability tests
- Heuristic analysis (great article here by Peep Laja, don't pay attention to the title, it's great even if you do have lots of traffic)



[...] an underlying theory
Put yourself in the shoes of your customers. Why didn't they do what you wanted them to? What was their intent at this stage? What would make YOU leave/not do what was expected?

Example: *"I think that if I'm a customer, I'd be more inclined to go through with this step if the form had less fields".*

Don't skip this. When you're deep in your CRO practice, you sometimes tend to forget about your users. You focus on numbers and design.
It's paramount to take a step back. The difference between a bad hypothesis and a great one could be staring at you in the face.

Let's not forget that we are beings doted with empathy. We're humans creating value for humans.

[...] an informed hypothesis [...]
With data and a theory, you can now craft your test hypothesis.

You can use this format or something along those lines:

By **{making this change}**, **{KPI A, B, and thus primary metric}** will improve **{slightly / noticeably / greatly}** because **{reasons (data, theory, ...)}**.

Working this way will not only improve the quality of your tests, but also your global conversion optimization efforts.

Good, you are now testing to confirm or disprove a hypothesis, not shooting from the hip hoping to hit something eventually.

3. They don't have a process and a roadmap

If you want to succeed at A/B Testing, and more largely at Conversion Rate Optimization, you need 2 essential elements:

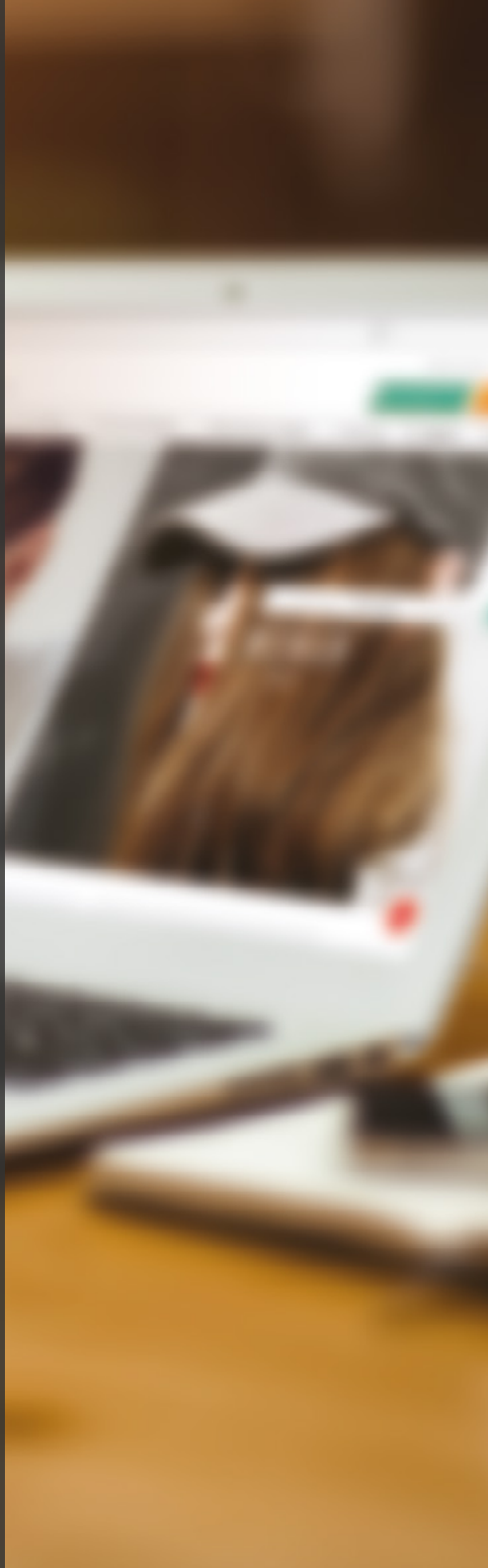
1. a process,
2. a roadmap.

1. Why do you need a roadmap? What does it consist of?

A roadmap will help you test what matters and work toward a clear end goal. It'll guide your efforts and prevent you from doing aimless testing.

- Business goals: the reasons you have a website. Be concise, simple, realistic.
- Website goals: how will you achieve these business goals through your site. What are your priorities?
- What are your most popular pages? Which ones have the highest potential for improvement?
- What does your conversion funnel look like, step by step? Where are the friction points?
- Key metrics: how will you measure success?
- A North Star: What's the one metric—correlated with Customer satisfaction, that if you focus exclusively your efforts on will guarantee your success? (Ex: Facebook=Daily Active Users, AirBnb= Nights Booked, Ebay=gross merchandise volume)

Put all these down and make sure everyone in your team/company is on the same page and can get behind them.



2. Sometimes when we hear "process", we have an allergic reaction. We think it means the death of creativity, that it's boring. You'd feel trapped.

It's absolutely not true for A/B Testing (or CRO). On the contrary. **It's BECAUSE you have a process that you'll be able to be creative.**

A/B Testing is a science experiment. You need to be rigorous, have a set of parameters making sure what you're doing is measurable, repeatable and accurate.

It's easy to fail. So you need safeguards. Steps that you will go through each and every time without having to think. No need to reinvent the wheel every time.

You need a process so you can focus on crafting the best hypothesis possible and optimize learning. The details of your process will be specific to you and your company.

But it will look something like that:

1. Measure
2. Prioritize
3. Test
4. Learn
5. Communicate (Don't skip this. Share with your colleagues why you did that test, how it is pertinent to your business goals, if it succeeded/ failed, what you learned. Encourage discussions, each department has insights for different stages of your funnel. And keep your IT team in the loop.)
6. Repeat

When you feel like you lost your focus, use Brian Balfour's question:

> *What is the highest-impact thing I can work on right now, given my limited resources (whether that's people, time or money)?*

Brian Balfour, VP of Growth at Hubspot

4. They don't prioritize

With a process, a roadmap and the development of a testing culture inside your company, you'll have a list of test ideas longer than your arm.

You need a system to prioritize tests, to stay on top of things and make sure you do what matters. And to not waste time and money on stupid tests.

There are articles on the interwebz roaming about and claiming things like: "CHANGING OUR CTA TO RED INCREASED CONVERSIONS BY 287%".

For the sake of argument, let's say they did have this crazy lift.

First thing first:

> There are NO colors converting more than others

What's important is visual hierarchy, i.e. is your CTA standing out from the rest of your page for example.

Second, testing minute details like color won't get you anywhere.

Most often, these type of changes are obvious. If your CTA doesn't stand out from your page, or if your headline/ copy isn't clear—do something about it. You don't need to test.

If you'd put it in the "Well, duh" category, don't invest your time and traffic in it. Just do it!

We're not saying you shouldn't test small things like adding a line of copy or changing the wording on a CTA. Just don't make them a priority (except of course—as we talked about earlier, if you're just starting out with A/B Testing).
If your ship was leaking, you wouldn't put a bandaid on a small hole in the deck.
Also, keep in mind that testing small changes or on low traffic pages equals small results.

It's true you should test everything, but you're working with a finite amount of ressources.

Once you're confident in your ability to A/B Test, focus on tests with the most impact/potential. Target your pages with the most traffic and influence.

You can use something like the PIE Framework to rate your tests so you know which one to do first.

Each test will be rated following three criteria:

1. Potential gain (../10): How much room for improvement is there on this(these) page(s)?
2. Importance (../10): How valuable is the traffic on this (these) page(s)?
3. Ease of implementation (../10): How easy will this test be to implement on your site?

Average the three for each test, and you'll have a backlog of ideas ranked objectively.

5. They don't optimize for the right KPIs

There are two types of conversions. Micro and macro. You should measure both, and optimize for the macro conversions.

A **micro conversion** is a step (click, social share, newsletter subscription, add to cart, ...) on the path leading to a **macro conversion**, which is an outcome that impacts your bottom-line (check out, free trial, ...), in other words, the main conversion goals of your website.

Why is it important that you know the difference?

Because you need to make sure that you measure both for every test, but that you don't optimize for micro conversions only.

There are two types of micro conversions **according to the Nielsen Norman Group:**

1. Process Milestones are conversions that represent linear movement toward a primary macro conversion. Monitoring these will help you define the steps where UX improvements are most needed.

2. Secondary Actions are not the primary goals of the site, but they are desirable actions that are indicators of potential future macro conversions.
Measuring micro conversions allows you to know where your friction points are and help you paint a holistic picture of your entire conversion funnel.

But you shouldn't optimize for them. You want to set your test goals as close to revenue possible.

You could get more traffic to your landing page through a given variation of your homepage, but have less form completions even though more people arrived on it.

So if you were optimizing only for the micro conversion of arriving on the landing page, you would lose money.

Track micro conversion, but don't optimize solely for them. And before starting any test, go back and make sure you're measuring everything that matters.

6. They ignore small lifts

"Go big or go home." It's true, as we've just said, that you should focus your tests on high impact changes first.

What you won't hear us say anywhere though, is "if your test results in a small gain, drop it and move on."

Why?

Because maths, that's why.

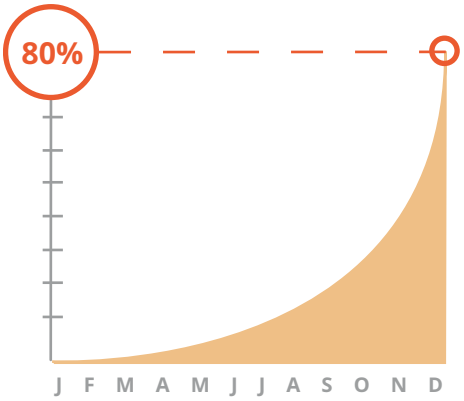
Let's take a quick example.

If a test gives your variation winning with 5% more conversions, and each month you get similar results. That's a 80% improvement over a year.

How's that for small!

Also, the more you test, the more you'll improve your website, the less you'll have big results. Don't be sad if you just get small lifts. It could mean your website is good. It's pretty rare to get big lifts on "normal" site.

> Don't be disheartened by small gains, it'll pay off big time over time.



7. They're not testing at all times

Every day spent without an experiment running is wasted. Opportunities missed.

Oh, look! An opportuni-aah... It's gone.

Why are you testing? Because you want to learn more about your visitors, make informed decisions, get more customers and earn more money.

When wouldn't you want all that? Never? Yup, same here. So as they say, Always Be Testing.

Testing properly takes time, you better not lose any. Test, test, test! And when in doubt, test again!

Next chapter we'll study the concepts needed to make better decisions as to when you can stop your A/B Tests or not. So you don't pull the trigger too early and end up with fake results (yikes).

> *"There's a way to do it better—find it."*

Thomas Edison

2

YOU'RE STOPPING YOUR TESTS TOO EARLY

YOU'RE STOPPING YOUR TESTS TOO EARLY

This is—without a doubt, the most common, and one of the most potent A/B Testing mistake.

But the answer is really not intuitive.

We were going to write a chapter on "when to stop your a/b test for both main methods of A/B Testing, frequentist and Bayesian".

But, we encountered two problems.

First problem, most people doing A/B Testing don't know—and don't care, whether their tool uses frequentist or bayesian statistics.

Second problem, when you dig a bit into the different solutions, you find that no 2 softwares use exactly the same statistical method.

So, how coud we write something helpful?

Here is what we came up with. **We will do our best to answer the following question: What concepts do I need to understand not to stop my A/B Test too early?**

Thus, we will cover:

1. Significance level
2. Sample size
3. Duration
4. Variability of data

Note: *None of these elements is a stopping rule on its own, but having a better grasp of them will allow you to make better decisions.*

1. Significance level

Don't trust test results below a 95% significance level. But don't stop a test just because it reached this level.

When your A/B Testing tool tells you something along the lines of: "your variation has X % chances to beat the control", it's actually giving you the statistical significance level.

Another way to put it is: "there is 5% (1 in 20) chance that the result you see is completely random".

Or "there is 5% chance that the difference in conversion measured between your control and variation is imaginary".

You want at minimum 95%. Not less. Yes, 80% chance do sounds like a solid winner but that's not why you're testing. You don't want just a "winner". You want a statistically valid result. Your time & money are at stake, so let's not gamble!

From experience, it's not uncommon for a test to have a clear winner at 80% significance, and then it actually loses when you let it run properly.

Okay—so, if my tool tells me that my variation has 95% chance to beat the original, I'm golden right?

Well ... no.

> Statistical significance doesn't imply statistical validity and isn't a stopping rule on its own.

If you did a fake test, with the same version of a page, an A/A test, you have more than 70% chance that your test will reach 95% significance level at some point.

2. Sample size

You need a sample representative of you overall audience (ignore that if you want to target a segment in particular for your test) and large enough not to be vulnerable to the natural variability of the data.

When you do A/B Testing you can't measure your "true conversion rate".

You arbitrarily choose a portion of your audience with the assumption that the behavior of the selected visitors will correlate with what would have happened with your entire audience.

You need to really know your audience.

Conduct a thorough analysis of your visitors before launching your A/B Tests.

Here are a couple of examples of things you need to know:

- How much of my visitors come from PPC
- Direct traffic
- Organic traffic
- Email
- Returning visitors
- ...

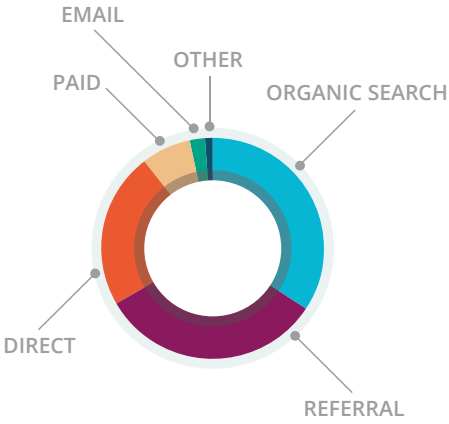
The problem is your traffic keeps evolving, so you can't know everything at 100% certainty.

So, what you need to ask yourself is:

How do I determine if my sample is representative of my entire audience, in proportions and composition?

Another issue if your sample is too small is the impact your outliers will have on your experiment.

> The smaller your sample is, the higher the variations between measures will be.



What does that mean? **Let's try with a real-life example.** Here's the data from tossing a coin 10 times. H (head), T (tail). We know the "true" probability of our coin is 50%.

1. We repeat the toss 5 times, and track the % of heads.

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	%H
T	H	T	H	T	H	T	H	H	T	50
T	H	T	T	T	H	T	H	T	T	30
H	T	H	T	T	H	H	H	H	H	70
H	T	H	T	T	T	T	H	T	H	40
H	H	H	T	H	T	H	H	H	H	80

The outcomes vary from **30% to 80%.**

2. Same experience, but we toss the coin 100 times instead of 10.

Tosses	%H
100	50
100	49
100	50
100	54
100	47

The outcomes vary from **47% to 54%.**

The larger your sample size is, the closer your result gets to the "true" value.

It's so much easier to grasp with an actual example.

With conversion rates, you could have your variation winning by far the first day because you had just shot your newsletter and the majority of your traffic were your clients for example.

They like you considerably more than normal visitors, so they reacted positively to your experiment.

Should you stop the test here, even with a significance level at 95%, you would have skewed results.

The real result could be the exact opposite for all you know. You made a business decision based on false data. Woops ...

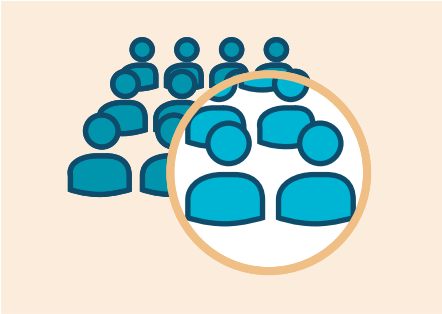
How big should your sample be?

There is no magical number that will solve all your problems, sorry. **It comes down to how much of an improvement you want to be able to detect.** The bigger of a lift you want to detect, the smaller sample size you'll need.

And even if you have Google-like traffic, it isn't a stopping condition on its own. We'll see that next.

One thing is true for all methods though: **the more data you collect, the more accurate or "trustworthy" your results will be.**

But it varies depending on the method your tool uses.



Here is what we tell our clients that use frequentist statistics.

Let us insist on the fact that those numbers might not be optimal for you if your tool doesn't use frequentist statistics.

That said, we advise our clients to use a calculator like this one (we have one in our solution, but this one is extremely good too).

It gives an easy-to-read number, without you having to worry about the math too much. And it prevents you from being tempted to stop your test prematurely, as you'll know that till this sample size is reached, you shouldn't even look at your data.

You'll need to input the current conversion rate of your page, the minimum lift you want to track (i.e. what is the minimum improvement you'd be happy with).

We then recommend at least 300 macro conversions (meaning your primary goal) per variation before even considering stopping the test.

I'll repeat it again, It's not a magic number.

We sometimes shoot for 1000 conversions if our client's traffic allows to. The larger the better, as we saw earlier. It also could be a bit less if there is a considerable difference between the conversion rates of your control and variation.
Okay, so if I have lots of traffic and a large enough sample size with 95% in 3 days it's great right?

Welp, kudos on your traffic, but sorry no again ...

3. Duration

You should run your tests for full weeks at a time and we recommend you test for at least 2–3 weeks. If you can, make the duration coincide with the length of 1 (or 2) business cycle.

Why?

You already know that for emails and social media, there are optimal days (even hours) to post.

People behave differently on given days and are influenced by a number of external events.

Well, same thing for your conversion rates. Don't believe me, try. Run a conversion by day for a week, you'll see how much it can vary from one day to another.

This means, if you started a test on a Thursday, end it on a Thursday. (We're not saying you should test for just one week.)

Test for at least 2-3 weeks. More would be better though. 1 to 2 business cycles would be great. You'll get people that just heard of you, people close to buying while accounting for most external factors (we'll talk more about those in our next chapter) and sources of traffic.

If you must extend the duration of your test, extend it by a full week.

4. Variability of data

> If your significance level, the conversion rates of your variations are still fluctuating considerably, let your test running.

Two phenomenons to consider here:

The novelty effect: When people react to your change just because it's new. It will fade with time.
Regression to the mean: This is what we talked about earlier: the more you record data, the more you approach the "true value". This is why your tests fluctuate so much at first, you have few measures so outliers have a considerable impact.

This is also why the significance level isn't enough on its own. During a test, you'll most likely reach several times 95% before you can actually stop your test.

Wait till your significance curve flattens out before calling it.

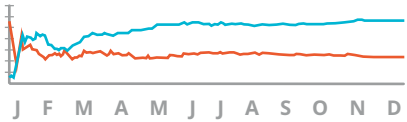
Same thing with the conversion rates of your variations, wait till the fluctuations are negligible considering the situation and your current rates. Some tools give you a **confidence interval**, for example "variation A has a conversion rate of 18,4% ± 1,2% and Variation B 14,7% ± 0,8%".

Meaning the conversion rate of the variation A is between (18,4 - 1,2) and (18,4 + 1,2), and the conversion rate of the variation B is between (14,7 - 0,8) and (14,7 + 0,8).

If the 2 intervals overlap, keep testing. Your confidence intervals will get more precise as you gather more data.

So, **whatever you do, don't report on a test before it's actually over.** To resist the tentation to stop a test, it's often best not to peek at the results before the end. If you're unsure, it's better to let it run a bit longer.

Next up, we dive, as promised into external validity threats.



3

WARNING!
IS THE WORLD SABOTAGING YOUR A/B TESTS?

WARNING! IS THE WORLD SABOTAGING YOUR A/B TESTS ?

A/B Testing is great, but let's be honest here: if you don't pay attention, it's pretty easy to fail.

You have to create an entire process, learn how to make hypothesis, analyze data, know when it's safe to stop a test or not, understand a bit of statistics ...

And... **Did you know, that even if you do all of the above perfectly, you might still get imaginary results?**

Did you raise an eyebrow? Are you slightly worried? Good. Let's look into it together.

What (or who) in the world is screwing up your A/B Tests when you execute them correctly?

The World is. Well, actually what are called **validity threats.** And you can't ignore them.

Your A/B Tests results might be flawed if:

1. You don't send all test data to your analytics tool
2. You don't make sure your sample is representative of your overall traffic
3. You run too many tests at the same time
4. You don't know about the Flicker Effect
5. You run tests for too long
6. You don't pay attention to real-world events
7. You don't check browser/device compatibility

1. You don't send all tests data to your Analytics tool

> Always have a secondary way to measure and compare your KPI's.

No tool is perfect, and humans are even less likely to be.

So, implementation issues with your A/B Testing tool are veeery common.

Your goals and metrics might not be recording properly, your targeted sample could be badly set up. Just to mention a couple of possible issues.

If you see any discrepancies in your data, or if a metric is not tracked correctly, stop everything.
Check your goals, metrics, scripts and start over. Same thing at the end of a test, make sure to double-check your results with your analytics tool before telling anyone.

Cross-check your data like your business depends on it (and it does).

It would be a shame to do everything right to have flawed results in the end because your tool was set up wrong or didn't work properly.

2. You don't make sure your sample is representative of your overall traffic

> You can't measure your "true" conversion rate, because it's an ever-moving target, new people come and go everyday.

When you do A/B Testing, you take a sample of your overall traffic and expose these selected visitors to your experiment.

You then assimilate the results as representative of your overall traffic and the conversion rate measured sufficiently close to what would the "true" value be.

This is why you should have traffic from all sources. New and returning visitors, social, mobile, email, etc in your sample, mirroring what you currently have in your regular traffic.

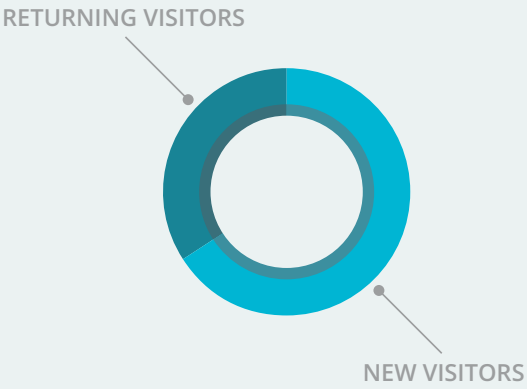
Unless you're targeting a specific source of traffic for your A/B test that is of course.

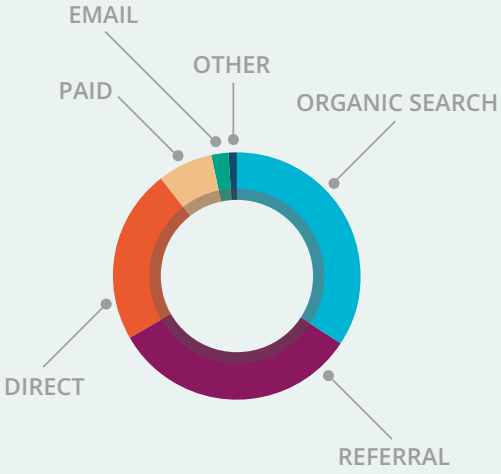
But if that's not the case, it's important you make sure you or your marketing don't have any PPC campaigns, Newsletters, other marketing campaigns or anything that could bring unusual traffic to your website during your experiment, thus skewing the results.
People coming to your website through PPC campaigns will generally have a conversion rate lower than regular visitors.

Whereas returning visitors—or worse, your subscribers, will be much more likely to convert. Because they already know you, trust you and maybe love you (I hope you have some that do love you. As Paul Graham, Co-founder of Y Combinator said: "**It's better to have 100 people that love you than a million people that just sort of like you**").

One other thing to keep in mind is: if your competitors are running some kind of big campaign, it could impact your traffic as well.

So pay close attention to your sample, make sure it's not polluted in any way.


RETURNING VISITORS
NEW VISITORS


EMAIL
OTHER
PAID
ORGANIC SEARCH
DIRECT
REFERRAL

3. You run too many tests at the same time

You're maybe familiar with Sean Ellis' High Tempo Testing framework. Or you just have enough traffic to be able to run several tests at the same time (well done you!).

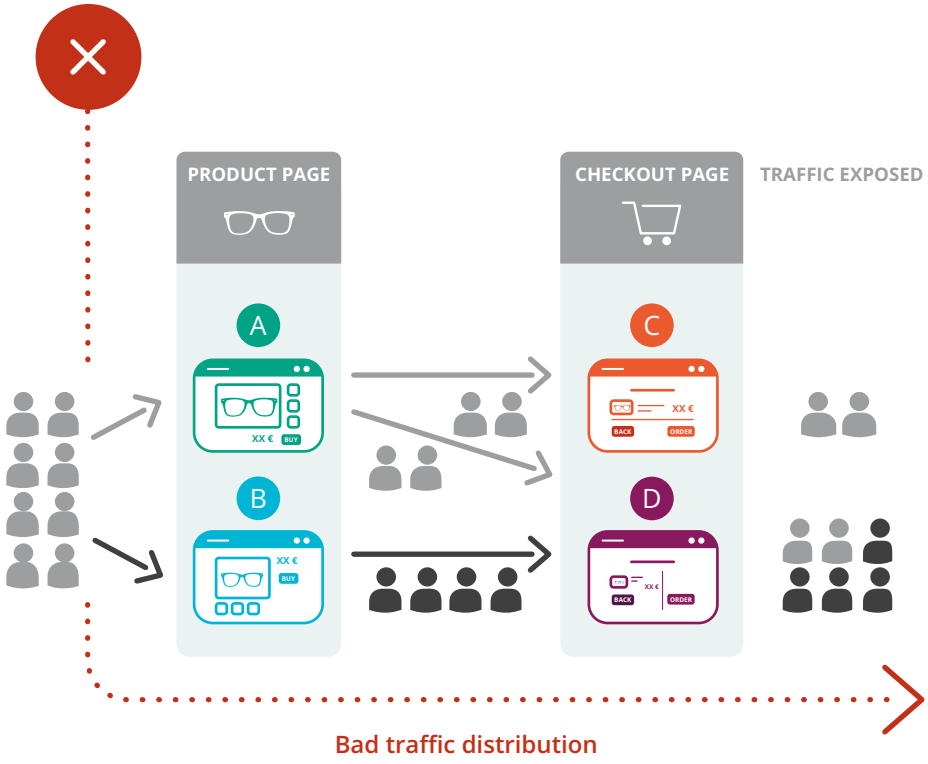
BUT—by doing that, you're increasing the difficulty of the whole process.

It takes time to gather data, measure, analyze and learn. So if you run 20 tests in parallel, I hope you have an army at your command. Oh, and that all your tests actually increase conversions.
Let's not forget that each test has a chance to DECREASE conversions. Yup. (*shudder*)

You might also be messing up your traffic distribution. My what now?

Let's take an example

Say you want to test steps 1, 2 and 4 of your check out process.

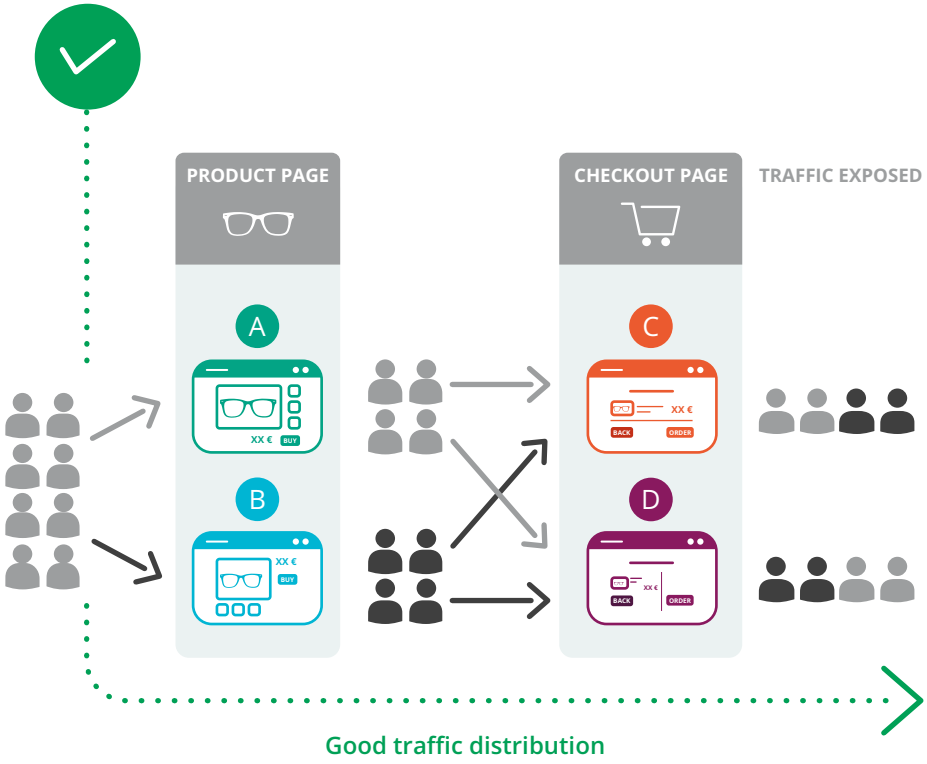


You send 50% of your traffic to your control step 1, and 50% to your variation. So far so good.
Except that you're also testing step 2.

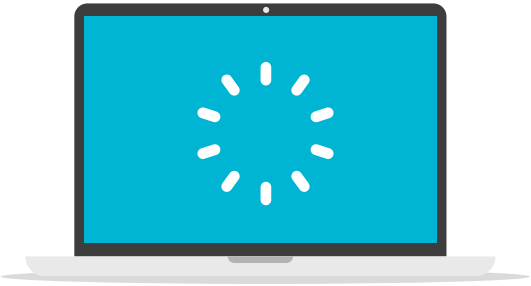
Meaning that if you don't re-split evenly (and randomly) people coming out of both your control and variation from step 1 so you have an equal amount of visitors exposed to your experiment, your step 2 experiment will be flawed. And I don't even talk about the step 4 one.

You absolutely can do multiple tests in parallel, but pay attention to traffic distribution, the influence tests could have on each other and that you can indeed handle all these tests properly.

You're testing to learn, so don't test for the sake of testing. Take full advantage of the learnings.



4. You don't know about the Flicker effect



> The Flicker effect (sounds like a movie title right?) is when people catch a glimpse of your control while the variation is loading.

It happens when you use a client-side tool because of the time needed for the JavaScript engine to process the page. It shouldn't be noticeable to the naked eye (i.e. less than 0,0001 second).

You don't want people seeing both versions and wondering what the hell is going on.

If it is noticeable, here are possible reasons:

1. Your website is slow to load (it's bad both for UX and SEO by the way)
2. Too many scripts load before your A/B Testing tool's
3. Something in your test is in conflict with / disables the tool's script
4. You didn't put the script in the <head> of your page

Optimize for the above reasons, or do split URL testing. Just make sure your control is not noticeable to the human eye.

If you want to know more about the Flicker Effect and how to fix it, check out this great article by Alhan Keser, Director of Optimization at WiderFunnel.

5. You run tests for too long

"Don't stop too early, don't test for too long" ... Give me a break, right?

Why testing for too long is a problem?

Cookies. No, not the chocolate ones (those are never a problem, hmm cookies).

Cookies are small pieces of data sent by the websites you visit and stored in your browser. It allows websites to track what you're doing as well as other navigational informations.

A/B Testing tools use cookies to track the experiments.

Cookies have a validity period (at Kameleoon the duration is customizable but they last 30 days by default).

In our case, if you let a test run for more than 30 days with the default validity period, you'll have people exposed to your test several times. When the cookies expire, your tool can't distinguish if the visitor was already exposed to the experiment or not, thus polluting the data.

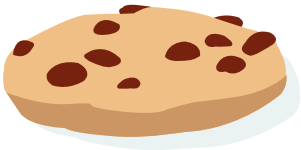
Considering that people sometimes delete cookies on their own too, you'll always have a degree of pollution in your data. You'll have to live with that.

But if you let your test run for too long, your cookies will expire. That's a real problem.

You could also get penalized by Google if you run a test longer than they expect a classic experiment to last.

> So make sure the validity period of your cookies is coherent with the length of your test beforehand.

And if you must extend (always by minimum a full week remember!) a test because you're not satisfied with your data, don't forget about them cookies.



6. You don't pay attention to real-word events

We already talked about how conversion rates vary between week days. **It's actually just one case among many other things, outside of your control, that can influence conversion rates and traffic.**

Here are a couple to consider:

Holidays:

Depending on the nature of your business, you could either have a traffic spike or drop, people having a sense of urgency and converting faster or the opposite.

Pay day:

Yup, it can be a factor in some cases. I mean, you can probably relate (I know I can), when it's pay day, you're way **more likely to buy certain types of products and make impulse buys.**

Major news:

Could be a scandal, assassination, plane crash, and it very well might be impacting your test. If people get scared, or if the news is sufficiently important for them to be distracted.

Weather:

First the obvious one: if a snow storm is coming and you sell parkas, you're going to sell more as soon as it's announced. But that's not all. Weather influences how we buy. It can be verified throughout History, and through more recent studies like this one by richrelevance that found for example a difference of 10–12% orders in cloudy vs sunny days for Clothing, Home/Furniture, and Wholesale retailers.

Time of day:

Depending if you're B2B, or B2C, people won't visit and buy at the same times. B2B will usually have more conversions during business hours, B2C outside business hours.

You can minimize the effects by keeping track of the news, looking at your annual data to look for unusual spikes in traffic, and see if they could be caused by external events.

7.You don't test for browser / device compatibility

This one is a make or break. Your test should work across all browsers & devices.

If variations don't work or display properly depending on device and browser, you'll end up with false results (as malfunctions will be counted as losses). Not to mention awful user experience.

Don't forget that 80% of internet users own a smartphone.

Be particularly careful with this if you use your A/B testing tool visual editor to make changes more complicated than changing colors or copy (we advise you avoid doing that, and code yourself for this types of changes). The generated code is usually ugly and could be messing things up.

Test your code on all browsers and devices. Chrome inspecting tool is usually good enough to simulate all devices (or through your a/b testing preview feature, if it has one like ours). Don't use browser-specific CSS or Ajax. Be careful with too recent CSS (check in your analytics tool what browsers your audience is using).

Next, we'll make sure you're not misinterpreting your results.

4

ARE YOU MISINTERPRETING YOUR TESTS RESULTS?

ARE YOU MISINTERPRETING YOUR TESTS RESULTS?

Once you have a process, know how to formulate a hypothesis, set up your A/B tests and know when to press stop, you're all good, right? You're on top of your A/B Testing!

Nope.

Making sure you're correctly interpreting your results is just as important.

A/B Testing is about learning, and making informed decisions based on your results.

So let's make sure we don't mess that up!

In this chapter, we'll look into ways you're possibly misinterpreting your tests results:

1. You don't know about false positives
2. You're not checking your segments
3. You're testing too many variables at once
4. You're giving up on a test after it fails the first time

1. You don't know about false positives

Are you aware that there are actually 4 outcomes to an A/B Test?

What do you mean, it's either a win or a loss, no?

Nope, it can be:

- **False positive** (you detect a winner when there are none)
- **False negative** (you don't detect a winner when there is one)
- **No difference between A & B** (inconclusive)
- **Win** (either A or B converts more)

(If you're a bit hardcore and want to know more about this, check out hypothesis testing. It's the actual mathematical method used for (frequentist) A/B Testing.)

Why should you care? Because you could have been interpreting false positives as genuine wins. And invested money in it.

Take the well-known example from Google: **41 shades of blue**. (Nope, it has nothing to do with the books. What books? No idea, you?)

Doug Bowman, Lead designer at the time, actually left (but for design reasons) the company because of this:

"Yes, it's true that a team at Google couldn't decide between two blues, so they're testing 41 shades between each blue to see which one performs better. I had a recent debate over whether a border should be 3, 4, or 5 pixels wide, and was asked to prove my case. I can't operate in an environment like that. I've grown tired of debating such minuscule design decisions..."
(You can read the full article here).

Whether you agree or not with him that it's wrong from a design standpoint, it's also mathematically wrong depending on how you do it.

You have two ways of approaching this:

1. You do "cascade testing", i.e. A vs B, then B vs C, then C vs D, ... THIS IS BAD, DON'T DO IT. We'll see why in a second.
2. You do A/B/n testing, meaning you test all variations in parallel.

1. Cascade Testing

Imagine you want to test a different headline for a product page. You have your current one (A) against the new (B). B wins, but your boss doesn't like the wording and wants you to try a slightly different wording. Then you feel like you could do better and change it again. And again.

You end up testing 10 different variations of this headline. How is it a problem?

Let's take a look: A vs B gave B as the winner with 95% statistical significance. As we saw in a previous chapter, it means that there is a 5% chance this result is a complete fluke or a "false positive".

Then you tested a third headline, B vs C. C also won with 95% significance. The problem is that the chance of a false positive compounds with the previous test. Your second test winner, C, has actually 9% chance of being a false positive.

After 10 tests on your headline (C vs D, D vs E, ...), even with 95% significance on your tenth test, you actually have a **40% chance of your winner being a false positive!** (For 41 variations it becomes 88%!!!)

You'd be flipping a coin. Or deliberately shooting yourself in the foot depending how many times you repeat this.

Don't do cascade testing. Just don't. Okay? Kittens will die if you do.

2. A/B/n Testing

A/B/n Testing is when you test n number of variations instead of just one (B) against your control (A). Meaning you have your control A, against variation B, C, D, E, F, etc. at the same time, in the same conditions.

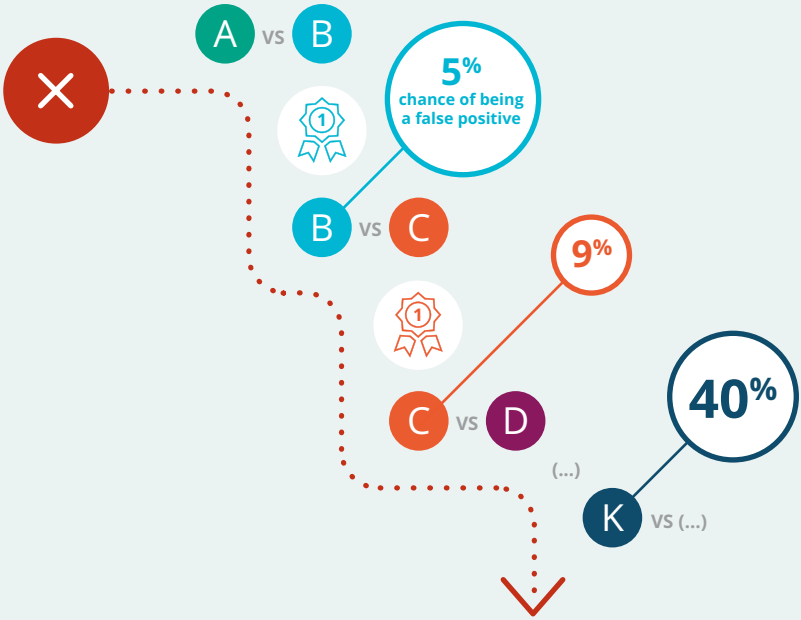
This is absolutely fine. BUT, as we saw in our chapter on when to stop your A/B tests, you need at least 300 conversions PER variation to call the test off if you're using a frequentist tool for example.

In our Google example, you would need 41x300 = 12300 conversions. That's a lot.

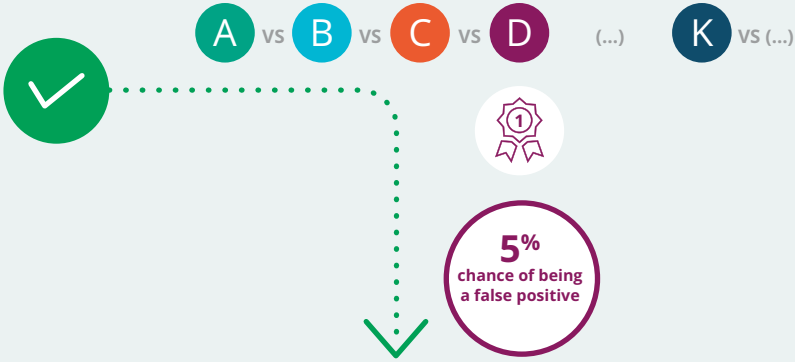
If you have Google-like traffic it's okay. If you're like us mere mortals though, this is a big fat loss of time.

You could even be testing for too long and get skewed results. **This kind of tests is very rarely needed and can often be completely avoided by having a better hypothesis.**

Cascade Testing



A/B/n Testing



2. You're not checking your segments

Don't make Avinash Kaushik sad (one of Web Analytics' daddies if you're wondering).

He has a rule:

> *Never report a metric without segmenting it to give deep insights into what that metric is really hiding behind it.*

Most data you get from your analytics tool is aggregated data. It takes all traffic and mashes it out into pretty but absolutely not actionable graphics.

Your website has a number of functions, your visitors come with different objectives in mind. And even when they come for the same reason, they probably don't need the same content.

If you want an effective website, you can't consider your traffic as a faceless blob, you need to segment.

It also applies for your tests results. If you don't segment them, you could be wrongly dismissing tests.

An experiment could be resulting in your variation losing overall but winning on a particular segment.

Be sure to check your segments before closing the book on a test!

Important side note: *when checking segments in an experiment's results, be sure not to forget that the same rules apply concerning statistical validity. Before declaring that your variation won on a particular segment, check you have enough conversions and a large enough sample size on that segment.*

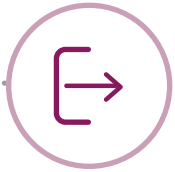
Here are three ways **you can segment your data:**

1. By source

Where do your visitors come from (Ads, Social Networks, Search Engines, Newsletter, ...)? Then you can look at things like: what pages they go depending on where they come from, their bounce rate, difference in loyalty, if they come back...

2. By behavior

What do they do on your website? People behave differently depending on their intent / needs.
You can ask: what content do people visiting your website 10+ times a month read vs those only coming twice? What page people looking at 5+ pages on a visit arrived on vs people who just looked at one? Do they look at the same products / price range?



3. By outcome

Segment by the actions people took on your website: bought a product, subscribed to a newsletter, downloaded a premium ressource, applied for a loyalty card, ...

Make groups of visitors with similar outcomes and ask the same type of questions we asked above.

You'll see what campaigns worked, what products to kill, etc...

By segmenting you get actionable data and accurate results. With actionable data and accurate results you can make informed decisions, and with informed decisions ... $$$$!

3. You're testing too many variables at once

You got the message, you need to test high-impact changes.

So you change the CTA, headline, add a video, a testimonial and change the text. Then you test it against your current page. And it wins.

Good, right?

Well... Not really.

How will you know which one(s) of your changes improved conversions on your page vs dragged it down?

This is where the question "How will I measure success" takes all its meaning. Testing is awesome, but if you can't really measure what happened, what moved the needle, it's not so useful.

What have you learned? That some combination of your changes improved conversions?

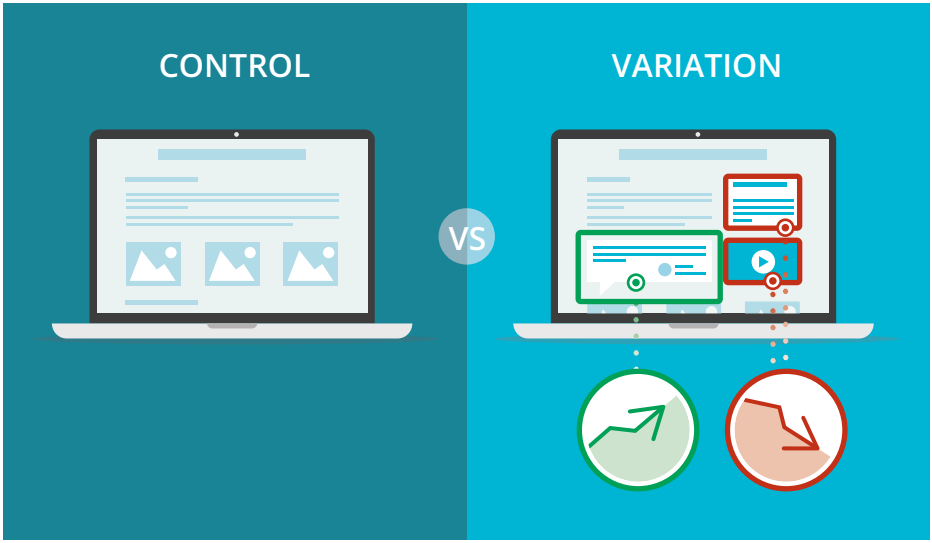
What if one of those positively impacted conversions and the others dragged it down? You counted the test as a failure and it wasn't one.

Make sure to clearly specify what success looks like and that you're set up to measure it.

If you can't measure it, you won't learn. If you don't learn, you can't reproduce nor improve it.

Don't test multiple variables at once. Unless you know how to do multivariate testing, then it's fine. But as it requires a gigantic amount of traffic, we rarely see it used.

Don’t test too many variables at once



How would you know which changes **IMPROVED** vs **DECREASED** conversions?

4. You give up on a test after it fails the first time

If you followed our guidelines on how to craft an informed hypothesis, each of your tests should be derived from (best should be a combination of several):

- Web Analytics
- Heatmaps
- Usability tests
- User interview
- Heuristic analysis

For example, you have analytics data showing people staying a while on your product page, then leaving.

You also have an on-page survey where visitors told you they weren't quite convinced that your product answered their need.

Your heuristic analysis showed you had clarity issues.

Click maps show people going through all your product pictures.

You then decide to test changing the copy and add better pictures on the page to make it more clear.

Your test ends and ... results are inconclusive, no increase in conversions.

What do you do now? You put a check mark in the "too bad" column, conclude clarity wasn't in fact an issue and move on to another test?

No, of course you don't. A/B Testing is an iterative process.

Take another look at your data, devise ways to improve your page.

- You could add testimonials
- You could remove information not relevant to the product
- You could add a video
- ...

As you now know not to do cascade testing (which is completely different than iterative because you don't test X versions of the same headline/picture against the winner of a previous test), or test everything at once, **you can embrace iterative testing.**

> There isn't just ONE solution to a given problem. There are an infinite number of them, and it could very well be a combination of several solutions.

> *If you're not stubborn, you'll give up on experiments too soon. And if you're not flexible, you'll pound your head against the wall and you won't see a different solution to a problem you're trying to solve.*

Jeff Bezos

Let's be a tad extreme to illustrate this.

When your internet cuts off, what do you do? If you're plugged through an ethernet cable, maybe you try unplugging/re-plugging it.

If it doesn't change anything, do you then conclude that your cable is dead, and go buy a new one?

Or rather you try to plug it in an another computer, go check your router, restart your computer, check your drivers, ...

Same thing with your A/B tests!

Don't give up or jump to conclusion as soon as something doesn't work. Look for other solutions and test again, and again.

Okay, you are now aware of several ways you could have been misinterpreting your A/B tests results, we're making progress!

For our last chapter, we'll take a look inside our brains, how they could be playing tricks on us and jeopardizing our A/B Tests *cue spooky music*.

5

YOUR BRAIN
IS YOUR WORST ENEMY WHEN A/B TESTING

YOUR BRAIN IS YOUR WORST ENEMY WHEN A/B TESTING

Did you know that we, humans, SUCK at statistical reasoning?

We're also irrational, flawed, and subjective.

Why? Because we're influenced by a list of cognitive biases longer than your arm.

You can perfectly live (but biased) without knowing about them, but if you're here, it means you're A/B Testing or contemplating to start so.

A/B Testing is a science experiment which must by definition be objective to provide actionable data.

Cognitive biases are then a real threat.

To get that out of the way, cognitive biases are personal opinions, beliefs, preferences that influence your ability to reason, remember, evaluating information.

Let's go down the rabbit-brain (sorry, had to do it) and make sure we're not subjectively influencing our tests too much by being our flawed (but lovable) selves.

Here's what your brain does and what we'll cover:

1. Will find a relationship between unrelated events
2. Can't handle sample size
3. Will subconsciously look for and interpret information to confirm its own beliefs
4. Will see patterns when there are none and thinking that past events influence future probabilities
5. Thinks what's in front of him is everything he needs to draw conclusions
6. Will base all subsequent thinking on the first piece of information received
7. Will throw reason out the window as soon as ego/emotion is involved
8. Prevents you from thinking like your customers
9. Takes things for granted because you've always done them this way
10. Overestimates the degree at which people agree with you

1. Finding relations between unrelated events

Remember how we talked about external validity threats?

Well, if you didn't know about them, you could assume that the lift you see was indeed caused by the pink CTA you put in your variation. Not because there is a storm coming that scared people into buying your product for example.

You'd have been victim of the **illusory correlation bias**. You perceived a relationship between 2 unrelated events.

Why an A/B Test worked or not isn't straightforward. Be careful not to rush your test analysis.

Our brain jumps to conclusions like there is no tomorrow. (A great book on the subject is "Thinking Fast and Slow", by Daniel Kahneman.)

> Your results are what you'll use to take business decisions, so don't rush your analysis.

2. Can’t handle sample size

When we talked about fixing a sample size before testing, we actually were also partially preventing another bias called **insensitivity to sample size.**

Our brain struggles to apprehend correctly sample size and underestimates variations in small samples.

Example from D. Kahneman’s book:

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

1. The larger hospital

2. The smaller hospital

3. About the same
(that is, within 5% of each other)

Care to make a guess?

Here’s what the answers looked like in the study:

“56% of subjects chose option 3, and 22% of subjects respectively chose options 1 or 2.

However, according to sampling theory the larger hospital is much more likely to report a sex ratio close to 50% on a given day than the smaller hospital which requires that the correct answer to the question is the smaller hospital."

Sample size is capital, and our brain usually forgets about it when considering problems similar to the example above.

Don't draw any conclusions from small samples as results won't have any statistical value. They can spark ideas, discussions and be the basis for actual tests, so you can verify the data.

You're warned of the risks now, so think hard before making a decision based of a small sample.

3. Looking for and Interpreting information to confirm your own beliefs

Confirmation bias is also not to be ignored. It's the fact that you will seek, interpret or focus on information (sub-consciously or not) that confirm your beliefs. Add to that the **Congruence bias**, where you test what YOU think is the problem, rather than the rest and you got yourself a nice, subjective test. For example, if you think the color red does indeed increase conversions, your brain will look for any information to confirm this belief.

A test could barely go in your direction (not in a statistically significant way), you'll be way more inclined to call it a success than if your convictions weren't on the line.

Every time you feel you were right, and that data seem to go your way, it's time to pause and ask yourself:

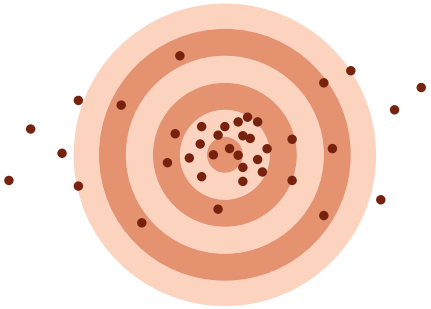
- Does it really prove your hypothesis in an objective way?
- Did you push this idea with your ego on the line to begin with?
- Aren't there other factors that could have produced (or at least considerably helped) this lift?

If you're testing to prove you're right and/or you value ideas themselves, you're doing it wrong.

> You're testing to learn, not building your ego. Impact is what's important.

4. Seeing patterns when there are none and thinking that past events influence future probabilities

Let's tackle two biases at once.



The **clustering illusion:** the intuition that random events which occur in clusters are not really random events.

A fun story illustrating the clustering illusion is the one about the Texas Sharpshooter.

It's the story of a Texan who shoots on the blank wall of his barn then draws a target centered where his shots are most clustered. And then he proceeds to brag about his shooting skills.

Because you see similarities doesn't mean there is a pattern. Nor because you made some good guesses in the past mean you'll keep making them. Flipping a coin 10 times and getting 7 tails doesn't necessarily means the coin is biased. It just means you got tails 7 time in a row

Okay, let's keep flipping coins. Let's say we flip another coin 39 times, and get 39 heads in a row.

What is the probability of having heads again for the 40th flip?

50%. Just as any other coin flipping.

If you were a bit confused, you fell prey to **the gambler** (or hot hand) **fallacy**.

You thought that because you got heads so many times in a row it would somehow influence the probability of the last throw.

Don't stop a test because "you see a pattern" or "spot a trend". Think about those 2 biases, odds are overwhelmingly in the favor of what you think you see is random.

Maybe you got some good results based on your intuition. Or so you think. You could actually be in the same position as our Texan shooter.

And because you've been right twice about something before, doesn't mean you will be next time.

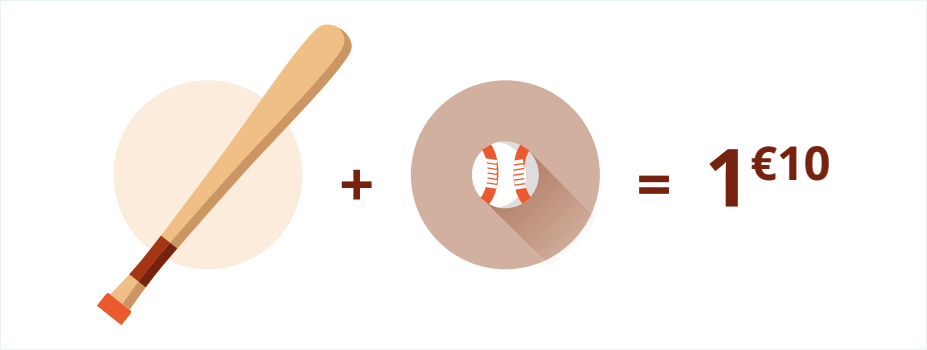
> Only data, obtained through rigorous tests, will tell.

5. Thinking what's in front of him is everything he needs to draw conclusions

This is what D. Kahneman called "**what you see is all there is**" in his book. It's the notion that we draw conclusions based on information available to us, i.e. in front of our eyes.

Doesn't sound too bad, uh?

Let's try with this:



A bat and a ball together cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?

50% of the students who were asked this simple question, students attending either Harvard or Yale, got this wrong.

80% of the students who were asked this question from other universities got it wrong.

I'll let you find out the answer on your own. And no it doesn't cost 0.10$.

Your brain is wired to look for patterns, and drawing conclusions with what you have. Except he sometimes jumps the gun.

Because you've got 2 pieces of data under your nose, doesn't mean they're all you need to draw a sensible conclusion.

6. Basing all subsequent thinking on the first piece of information received

Called **the Anchoring bias**, it's the fact that we allocate more importance to the first piece of information we're given.

Here is an example from a study by Fritz Strack and Thomas Mussweiler: 2 groups of people were asked about Gandhi's age when he died.

The first group was asked if it was before 9 years old or after. The second if it was before 140 or after. Both answers are pretty obvious.

But what was very interesting, were the answers from both groups when they were asked to guess Gandhi's actual age when he died.

Answers from the first group had an average of 50 vs 67 for the second. Why such a difference?
Because they were subconsciously influenced by their respective first questions.

Think about your last salary negotiation for a job interview. The first person to give a number basically calibrates the rest of the negotiation, because everything following will be based on that number, for the starting value and the scale of the negotiation.

When the number given is precise, people tend to negotiate in smaller increments than with round numbers (studies on the topic here).

If the interviewer went first, with his highest bid—or what he said was his highest bid, I'd wager you used it as a base with your counter-offer instead of what you thought you were worth.

By now you must be getting weirdly suspicious of your own brain. Good. Being aware that we're built to jump to conclusions, consider only what's in front of our eyes and ignore the big picture is the first step in the right direction.

Be extra-careful with your numbers and tests! When you feel you're sure about a result, pause and check again. Run a the test a second time if needed.

7. Throwing reason out the window when ego and emotion are involved

This one can be painful.

You put your heart and soul in a redesign, you spent hours on it and you're super-proud of what you made.

Then you test it. And it flops. Badly. Ouch ...

What do you do?

"Screw these people my design is perfect, they don't know what they're talking about!"

Or when you bring the news to your boss he says: "No way, this design is clearly better, go with this one"

No! Be strong, I know your pain. **It's hard but that's one of the reasons you're A/B Testing, to not lose money on redesigns or decisions based on guts and personal opinions.** But rather do things people actually want.

Swallow your frustration, go back to the hypothesis that led to this redesign, aaand to the drawing board again.

> Being able to throw out hours —days even, of work through the window if your data says so, is a sign you're truly becoming data-driven.

It's freakishly hard though.

8. Preventing you from thinking like your customers

Called the **curse of knowledge**, it's when you've been so absorbed by a subject that you're having a hard time thinking about problems like someone who has little to no knowledge about it.

When you know something is there—say a new button or a new picture, that's all you see on the page.

But your visitors could just as well not even see the difference.

Ask someone else, a colleague from another team to take a look. Or you could do usability tests.

Don't ask someone from your team though. You could all be victims of **the bandwagon effect.** Members of a group influence each other. And the more people do something, the more other people might be influenced to do the same.

If you don't regularly receive external feedback, you might have built yourself a distorted reality.

Regularly ask your visitors and clients, as well as other teams for feedback.

9. Taking things for granted because you've always done them this way

Functional fixedness is when you're stuck in linear thinking.

It means that because you've always done something in a certain way or used a tool a certain way, you don't question it and go with it.

You see an iron, you think about its obvious use—for clothes, you don't think to use it as a toaster if you don't have an oven (or a real toaster).

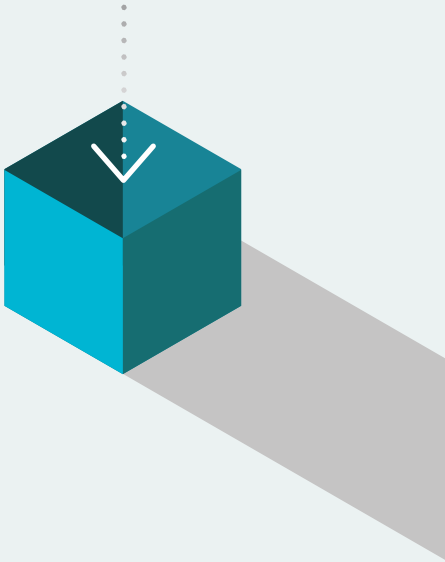
This is called **lateral thinking**, or "thinking out of the box". Easier said than done though, uh.

You can try to trigger this type of thinking by repeatedly asking "why" every time you think something is obvious. That way you'll dive in your assumption till you get to its bottom—or source.

What you'll find might blow mind, sideways.

Other things you can do to try and trigger lateral thinking:

- Turn your problem on its head, try to solve the opposite
- Think about the most obvious, stupid solution
- Fragment your problem in series of small, very precise problems
- Don't be satisfied with finding one solution
- Flip your perspective, how would you think about this problem if you were an engineer, a scientist, a complete beginner?



10. Overestimating the degree at which people agree with you

"Everyone hates popups."

Well, YOU hate them. They usually annoy people a bit but they actually convert quite nicely when used the right way (i.e they don't pop in your face the second you arrive on a website).

This type of bias (aka **the false consensus effect**) can be tricky when gathering feedback.

We can sometimes believe strongly in something, and think we're on the side of majority when we're really not. We tend to assume people have the same opinions we do. Particularly in a small group, when we arrive to a consensus on something, we're very quick to think it's representative of the overall population.

That's why it's better to get feedback from individuals rather than from a group. Group feedback will be plagued by biases.

But when asking individuals, don't get caught up in personal preferences either. Always take everything with a grain of salt.

Be careful not to do it yourself too!

When you think that something is doing as good as it could be, stop in your tracks and reconsider.

> Test what would have the most impact, but also test what is doing fine. You can always do better.

That's all on cognitive biases... and for our series on A/B Testing Mistakes.

Congratulations! You went through all 67 pages of this ebook! We hope you got some useful nuggets , learned a bit and that your tests will get better as a result.

Back to you now, let's go get some testing done, mistake free!

NEVER MISS OUR EPIC CONTENT

If you're serious about optimizing your website, we bet you'll love our monthly newsletter.

More ebooks like this one will come too.

SIGN UP FOR FREE

Conversion Matters
by Kameleoon

Kameleoon is the most advanced optimization platform on the market. Simple, fast and reliable, our SaaS A/B testing and personalization solutions have been designed to allow marketing teams to craft optimized and personalized experiences to each of their visitors and to take the right decisions at the right time.

Kameleoon's solutions rely on an architecture that treats real-time visitors data and runs predictive targeting algorithms able to identify optimal segments for personalization campaigns.

Several hundreds of clients, from pure players to large firms, have successfully adopted Kameleoon's solutions to boost their conversions.

FOLLOW US

 www.kameleoon.com

FRANCE
12, rue de la Chaussée d'Antin
75009 Paris, France
+33 (0) 1 83 62 20 50
info@kameleoon.com

GERMANY
Beim Alten Ausbesserungswerk 4
77654 Offenburg, Deutschland
+49 (0) 781/924 17-726
info@kameleoon.de

RUSSIA
105187 Россия, Москва,
ул. Щербаковская, 53/3
+7 910 413 20 83
info@kameleoon.com