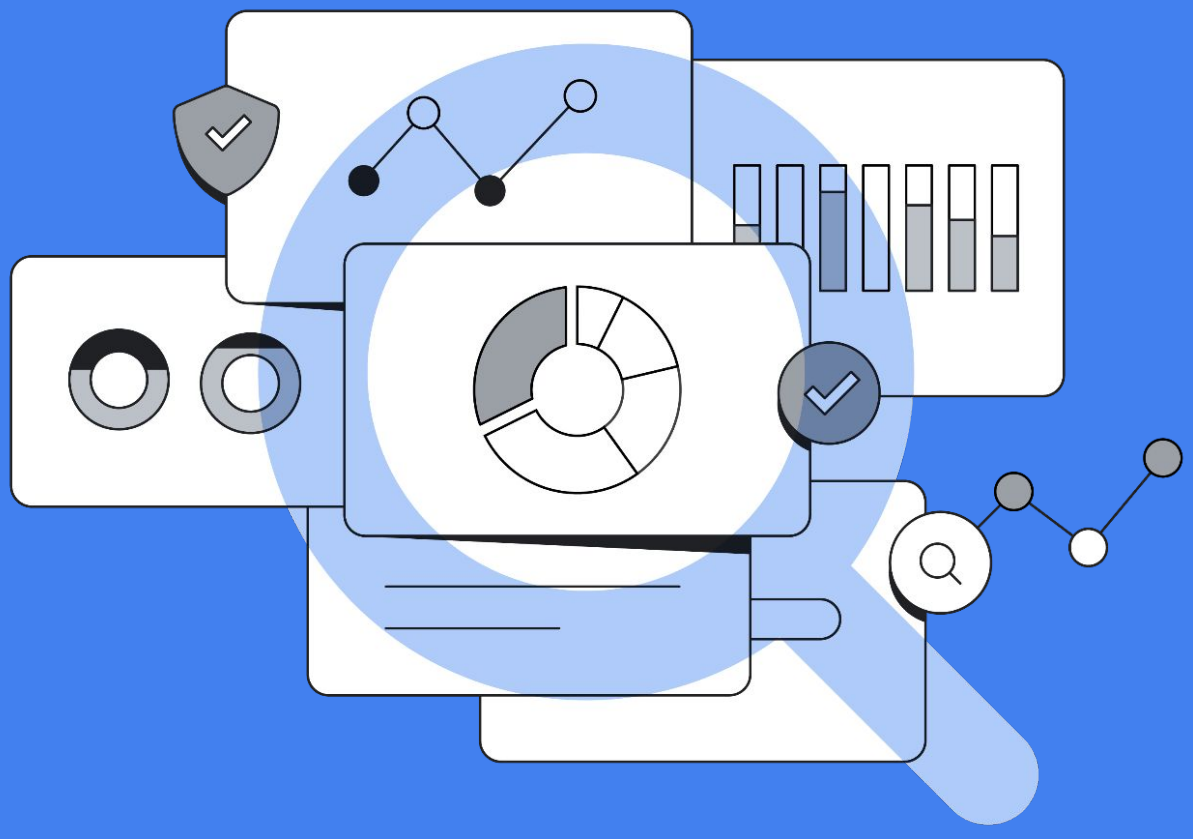
Google

The Experiments Playbook

How to set up *high-quality experiments* to **prove** and **improve** the effect of your media

2024 edition



Google

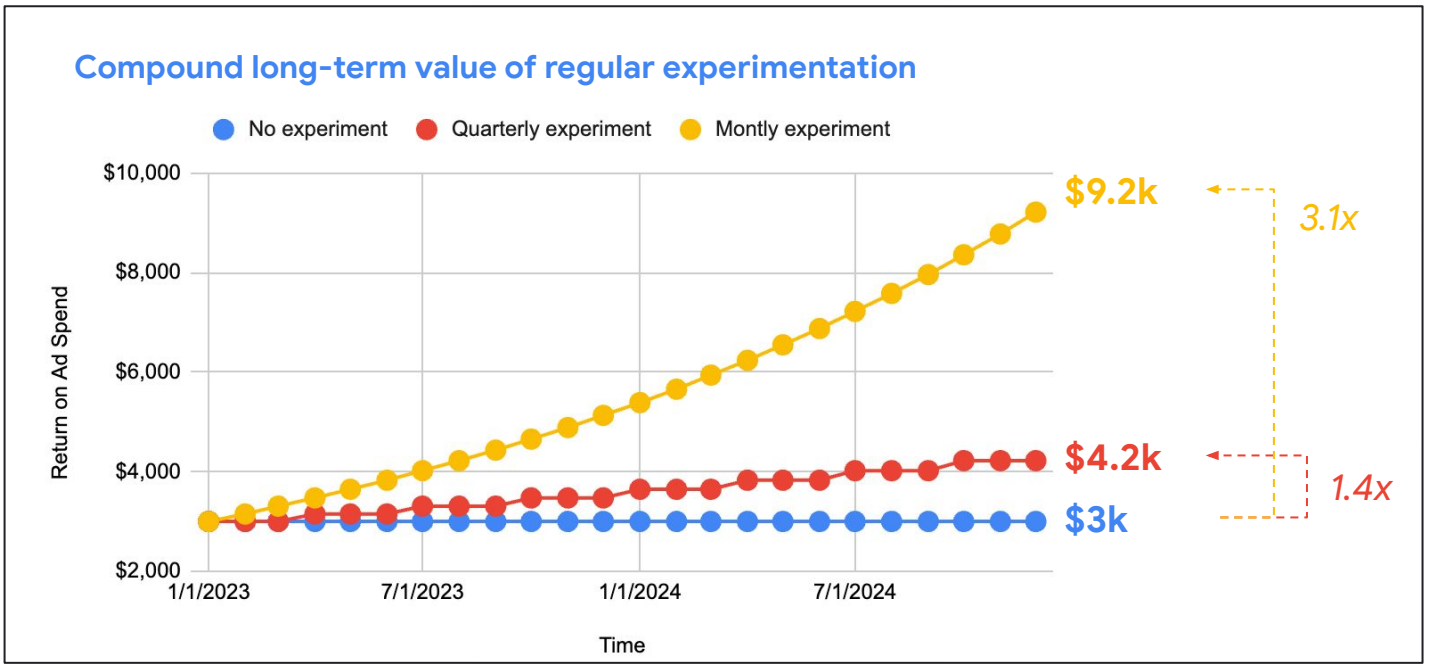
The value of experimentation

Marketing & advertising, if executed well, can serve as a **competitive advantage** for any company. It allows them to *grow faster* than their peers and *more effectively*. But how can we determine whether marketing is being executed well? That's where **experiments** come in.

Increasingly more advertisers are turning to experiments to answer fundamental questions of media effectiveness. Should I activate this new marketing channel? How many incremental conversions did this video campaign generate? Should I revamp my creatives? Who is my optimal target audience? How high should my budget be to increase conversions? Should I use tROAS or tCPA bidding?

Experiments are a powerful tool to answer these questions. They allow us to measure the **causal relationship between a treatment and an outcome**. They enable us to go further than just *correlation* and translate to *causality*. Not only can experiments be used to *prove* the value of media, they are used more and more to *improve* the performance of marketing. As the saying goes "*if you can not measure it, you can not improve it*".

By regular experimentation on marketing activities companies are able to build a **competitive advantage / measurement muscle** that allows them to grow more quickly than their peers.



The graph above illustrates the value of experimentation. The **blue line** represents the return on your marketing investment if you **don't perform experiments**. When no experiments are performed, we can't truly learn what works and what doesn't, so we can not improve marketing effectiveness. However, if **experiments are performed on a quarterly basis**, and we learn how to improve our marketing with 5% with every experiment, we see there is a **1.4x increase in returns on marketing investment** over the period of 2 years. In the situation of **monthly experiments**, assuming we learn how to improve our effectiveness with 5%, we see a **3.1x** increase in returns over the same time period.

However, with **great power comes great responsibility**. As the enthusiasm and demand for experiments is greater than ever, the need for guidelines on how to run high-quality experiments has reached new peaks. Incorrectly set up experiments cost a lot of *time*, *resources* and lead to *suboptimal actions*. In this guide, we will discuss the 4 main characteristics of valid and accurate experiments that can fuel your business for growth.

Google

Running high-quality experiments

The experimentation squad

Before diving deeper into what characteristics turn an experiment into a high-quality experiment, it is useful to discuss what types of people or knowledge we need to successfully run experiments in the long term. So, who do you need in your "**experimentation squad**"? We believe you need 4 sets of skills:

I believe we should improve our marketing this way!

The creative one
Come up with ideas for potential new experiments

That change could result in an 10% increase in revenues

The business person
Prioritise which experiment ideas are most impactful

An experiment to test this hypothesis looks like this...

The data genius
Ensure the statistical side, design, run & evaluate the experiment

I support you to test & if successful we can scale it globally!

The decision maker
Dedicate time & resources to run experiments and act upon them

What makes an experiment "*high-quality*"?

At Google, we believe that high-quality experiments have 4 important characteristics. More specifically they:

1. Link **questions** to **actions**
2. Are about **great metrics**
3. Consider the **statistics**
4. Are seen in their **context**

Even though this sounds simple and easy, there are multiple (subtle) **pitfalls** that are oftentimes observed when experiments are set up. We'll try to cover the most common pitfalls and how to make sure they are not present in your new experiment.

Happy experimenting!

Google

1. They link questions to actions

Start with a relevant business question

The first characteristic of high-quality experiments is that they start with a clear and relevant research question that is linked to the business objectives or most important organisational KPIs. A hypothesis is formed, based on people's opinions, thoughts or previous analyses that have been performed throughout the organisation. Based on this hypothesis, a research question is formulated.

Pitfall #1: *Across marketing teams, confusion tends to arise around what exactly we are trying to achieve in a certain experiment. The digital marketer might say we are optimising for revenues, while the data scientist might say we are optimising for clicks.*

To eliminate this confusion, it's a good practice to **document the hypothesis and / or research question formally in an experiment tracker**. By formalising hypotheses and research questions we create a sense of clarity and accountability, as everyone can access them and ask questions.

Pitfall #2: *Whenever a new experiment idea or research question is proposed, think critically about how answering this specific question will move the needle for your business. Does the experiment help answer a question that will influence an important decision to take the right action, at the right time?*

Experiments are valuable tools to answer questions, but they also **require lots of time, resources and data** to set up and evaluate the experiments. Consider the opportunity cost of all this time and resources that is needed for a given experiment and think about whether the value of knowing the answer to this question is balanced compared to the "*cost of experimentation*".

Link potential outcomes to actions

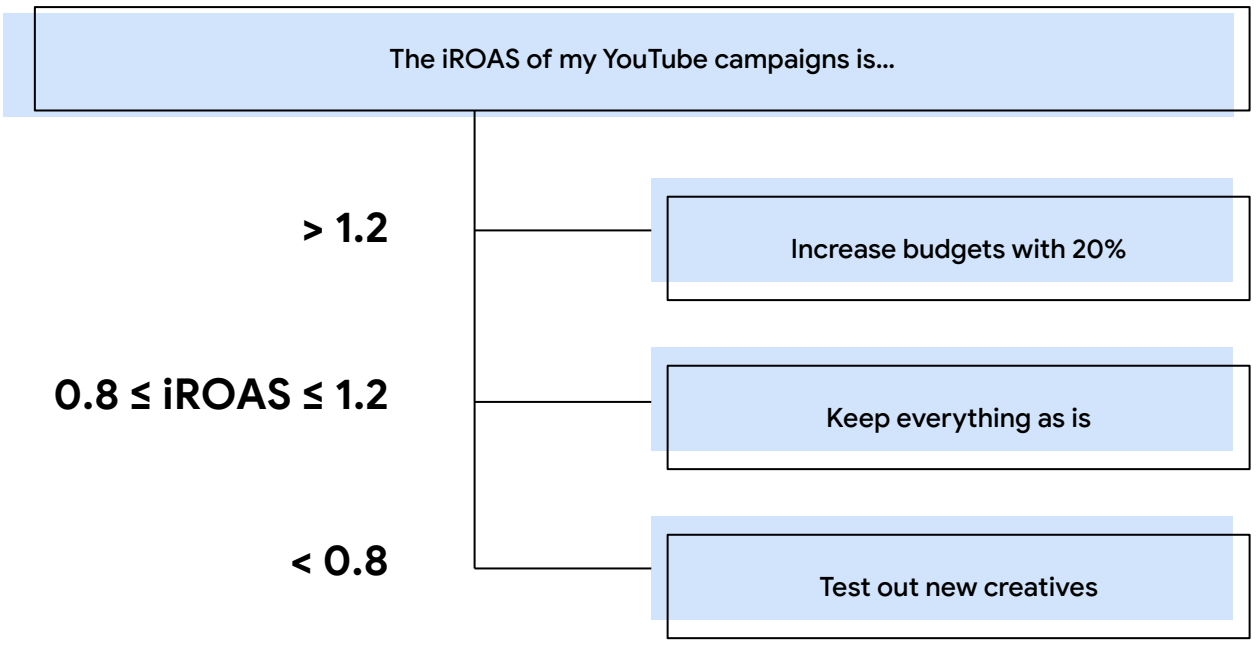
The point of running experiments is to learn something we couldn't know otherwise. This newly learned information should allow us to take actions that we would not be able to have taken with confidence before. Therefore, it is important to define the actionability of the experiment upfront. What are the actions that will be taken in each potential scenario of outcome? What will you do differently when the experiment is over?

Pitfall #1: *Experiments are set up to answer a question that can not be (feasibly) taken action upon. For example, let's say we set up an experiment to measure the differences in incrementality of our YouTube ads between young people and old people. If however, we cannot target based on age group in YouTube, setting up this whole experiment is not relevant as we can not take action upon our learnings.*

This pitfall is related to the idea of running experiments because "*we just want to know*". Even though knowing something is valuable, experiments take a long time and much resources to design, run and evaluate, so it is always a best practice to ensure actions based on experiment outcomes.

Pitfall #2: *Experiments do not result in any change of course of action, even if the research question we answered is actionable. That raises the question, why did we spend all this time and resources on this experiment?*

To overcome this pitfall, we advise **drafting an action tree** (see below), outlining what action will be taken in each potential outcome of the experiment. It is best practice to draft this action tree *before* actually running an experiment, during the experiment design phase, to make sure we are thinking about actionability when setting up the experiment. Also, **align with the decision makers** to agree what actions will be taken in each scenario.



Google

2. They are about great metrics

Important and relevant metrics

In high-quality experiments, the KPI in question is connected to the business and / or marketing objectives.

Pitfall #1: *It is not uncommon that marketing experiments are designed around optimizing for something that does not relate to actual business value, such as: research questions around clicks, impressions, CPMs, attributed conversions, attributed revenues, click-through-rates, ... without these being linked necessary to bottom line revenue or profit for the company.*

To solve for this, try to **link the objectives of the experiment to the broader business goals**. What are you as an organisation aiming for? How is this KPI contributing to that broader goal? For example, a mobile gaming company should not be optimising for CPMs but optimising for games played, which is directly related to revenue made. By making this transition, they can measure the impact of their actions and interventions on what truly matters for the business.

Framework tip: The effectiveness - efficiency KPI framework

Marketing is all about growth or generating more revenues (*effectiveness*), however, we also want to be aware of our costs (*efficiency*), to make sure we are not wasting money.

A framework we see great success with is the effectiveness - efficiency framework, where we define

- 1 goal KPI: iConversions, iRevenues, iAppInstalls
- 1 constraint KPI: iROAS, iCPA

The goal is to *maximise the goal KPI*, while *staying at or above the constraint KPI target*.
This way you ensure maximum growth for the growth cost you are willing to pay.

Google

Measurable metrics

When running experiments, we expect there to be random noise in the data. This is not a problem when we have a lot of data and the effect of the treatment is strong. However, if our KPI is too sparse (i.e. there is a low volume, infrequent), the noise can be too high for the experiment to be able to measure a significant lift. If that is the case, we would recommend to choose another metric that is less sparse, but correlated to your primary business KPI.

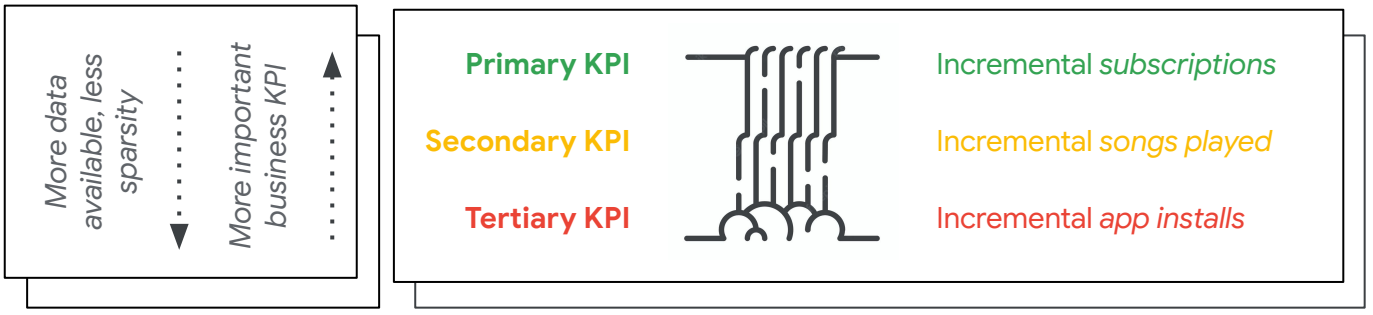
Pitfall #1: *A car company's most important metric is the revenue it makes from selling cars. However, car purchases are quite a sparse event, as people do not buy new cars regularly. Therefore, when running an experiment to measure the effect on car purchases, it will be very difficult to get significant results.*

To overcome this, we could look at higher-in-the-funnel metrics, such as car dealership visits, appointments made, Google searches or website visits as a metric as long as it correlates well with the primary KPI, revenues.

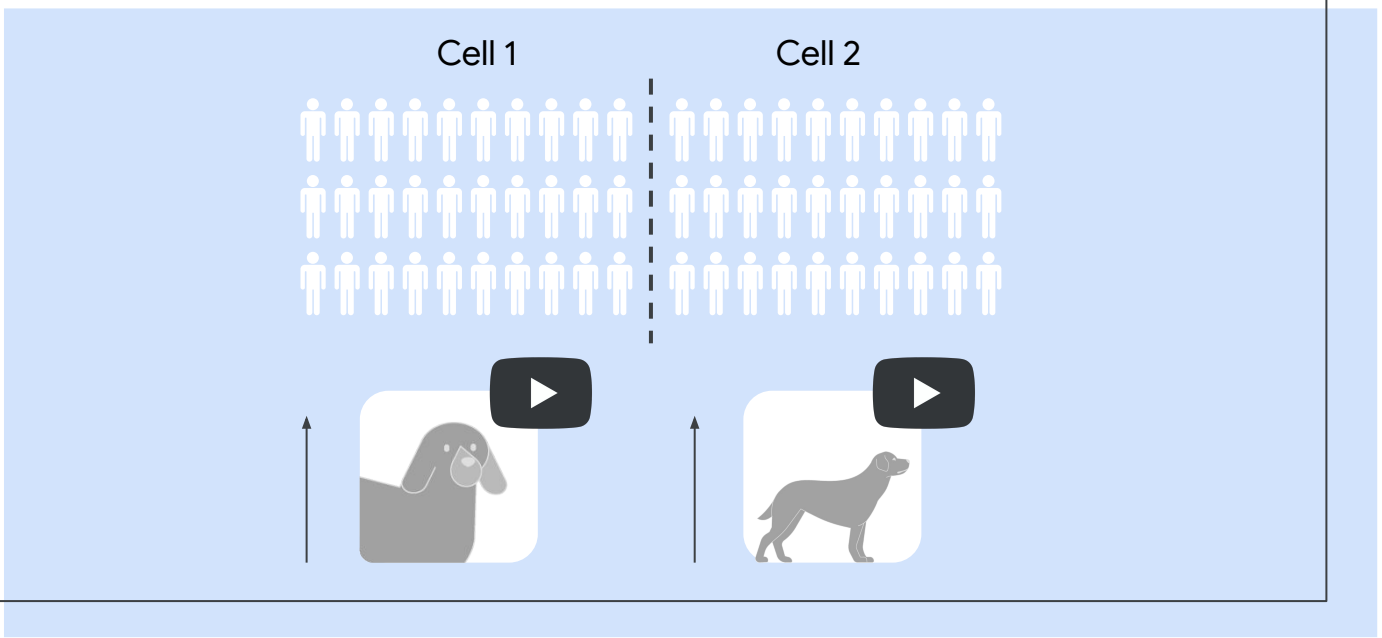
Oftentimes it can be useful to define a set of **secondary** or **tertiary metrics** that can serve as goal metrics in a waterfall framework. These metrics would only be used as diagnostic tools to support hypotheses or to make decisions when we can not make a decision only using the primary goal KPI.

A great secondary metric is one that is less sparse than the primary metric, but does correlate well with the outcome.

An example of the **KPI waterfall** for a music streaming company



To see how this would work in practice, let's consider the following situation. **ChewyTreats** is an online dog food seller and has been actively advertising on YouTube for years. However, new research from the Canine Research Institute has shown that animated video ads would be more effective in generating conversions compared to real-life footage, which they have been using this far. The creative and data science team are eager to set up an experiment, splitting up the total audience into 2 groups (cell 1 and cell 2, see next page), and measure the incremental effect of these different creative videos.



A couple of weeks later, when the experiment is completed, the team at **ChewyTreats** receive the following results in the report:

Metric	Cell 1	Cell 2
Incremental Conversions	*[12k - 16k]*	*[13k - 17k]*
Incremental Add To Carts	*[34k - 40k]*	*[39k - 45k]*
Incremental Visits	*[100k - 130k]*	*[140k - 165k]*

As the confidence intervals* for *Incremental Conversions* and *Incremental Add to Carts* are overlapping, we can not - with the desired level of confidence - make an assessment of which creative is performing better for these metrics. Directionally, we would say that Cell 2 is performing better than Cell 1, as the confidence intervals are directionally higher, but this can also be due to noise and random fluctuations in the experiment.

However, looking at a higher-in-the funnel event, such as *Incremental Visits*, we do see a significant difference, as the confidence intervals are not overlapping. We can therefore, with the desired confidence, make the statement that the animated video is delivering more *Incremental Visits* than the real-life footage video.

By applying this waterfall framework, the marketing team at **ChewyTreats** is able to make a decision on which creative format to move forward with, while if only results were recorded for Conversions and/or Add To Carts, we would not have been able to make a decision.

** A confidence interval is a range of values we believe with a certain level of confidence the actual value is between.*

Google

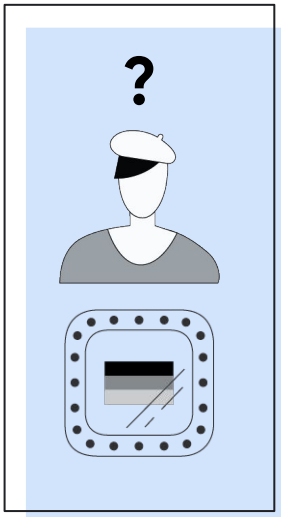
3. They consider the statistics

A fair comparison

Essential for a high-quality experiment is that we ensure that we compare a group of treated subjects with a *similar* group of non-treated subjects. This is called **internal validity** in statistics, and is imperative for the quality of the experiment.

Pitfall #1: *A large perfume brand in EMEA, wanted to know the impact of activating Performance Max campaigns on their KPI, sales. To answer this question, they activated a set of campaigns in France and did not activate these campaigns in Germany. After a couple of weeks, they start comparing the sales between France and Germany to see whether or not they could measure a lift, however they saw incredibly positive results...*

The perfume brand did not perform any analyses to ensure that the groups they are planning to compare are similar in terms of characteristics and how subjects would respond to the treatment (the PMax campaigns). In many cases it does not make sense to compare people from one country with another country, due to differences in *demographics*, *income levels*, *legislation* and *price setting*, as well as other differences in *marketing activations* and *competitive dynamics* and *brand power*. When doing geo-experiments, we therefore recommend to set up **within-country experiments** instead of **across-country experiments** and it is always a good idea to perform a **correlation analysis** when deciding on treatment / control assignment.



Conversion Lift studies inherently do not have issues with this problem, as the random split between treatment and control users guarantees similar groups (referred to as *homogeneity* in statistics).

Do the math

Experiments study human behavior, which inherently contains a lot of random noise. This means experiments always have to deal with uncertainty. Due to this uncertainty in the data, we have to make sure we collect enough evidence (i.e. data) to make decisions, considering a desired level of confidence. To ensure this, we consider the *power* for each possible experiment setup.

Pitfall #1: *Furniture company X runs a geo experiment for 3 weeks. After the experiment, they see the results are not significant, although a directionally positive impact is observed.*

During the experiment design phase, we should perform a **power analysis** which assesses the feasibility of a given design in terms of experiment duration. Based on the variation it sees in the data and an assumption around the size of the effect, the power analysis will tell us what is the minimum experiment duration to measure a significant effect.

** A power analysis is a feasibility study to ensure we can reach statistical significance under certain assumptions such as minimum detectable lift and amount of data*

Google

4. They are seen in their context

Take external variables into account

The world is constantly changing and we cannot always control what changes are affecting our metrics / KPIs. However, it is important in an experiment that we understand which external variables that are impacting our results, and whether or not they impact treatment and control in the same proportional manner.

Pitfall #1: *The local marketing team for a large shoe retailer launches a local promotion campaign in a group of cities. At the same time that we are doing a geo-experiment to measure the incrementality of a new video marketing channel. After investigation it appears that 9 out of 10 of these cities were all in the control group of the geo-experiment. The geo-experiment that was in planning for multiple months will need to be restarted due to this local campaign.*

Experimentation and testing should be approached from a **holistic perspective**. Ensure regular catch ups between teams, knowledge sharing and other processes to spread this information.

Do not generalise learnings

Regardless of how well-designed an experiment is, it is important not to forget: *What we measure in an experiment is only valid during that experiment and in the context of that experiment*. In statistics this is called **external validity**, so make sure we do not generalise learnings in ways we can't. Media effectiveness is a dynamic concept, just like almost every treatment effect & is often different across geographical areas.

Pitfall #1: *Sometimes we hear from advertisers something along the lines of "We did an incrementality test 1.5 years ago and we saw that our YouTube campaigns were 4% incremental. Since then the economic reality has changed dramatically as we seem to be heading towards a recession, also we have increased our search budgets by factor 2, we are doing a lot of Discovery and even have a huge sponsorship with Mickey Mouse, but we still assume the same incrementality".*

Regular testing and **experimentation** is recommended as the environment is always changing and will make sure we have an accurate view on marketing effectiveness so we can optimise even further.