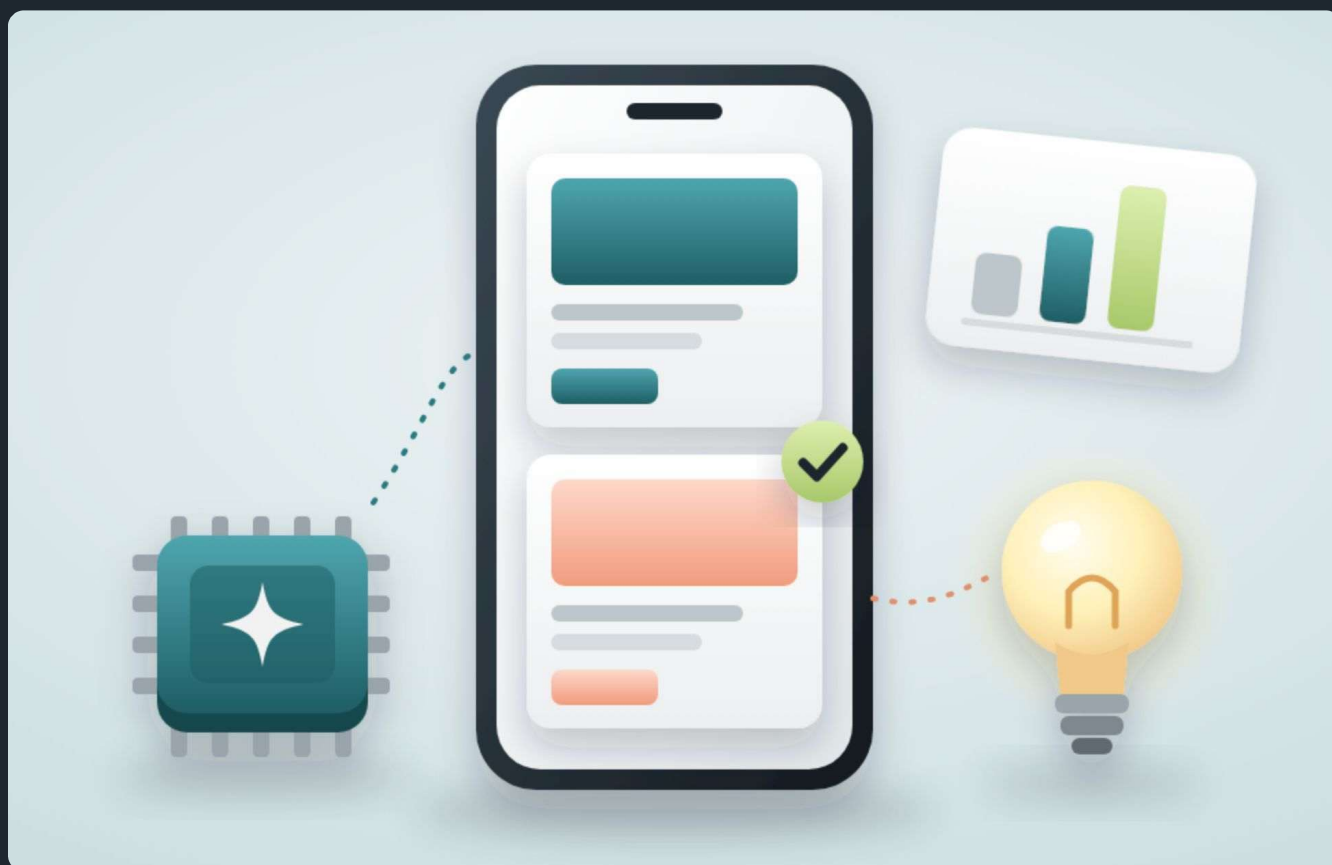




Why so many experiments are inconclusive

And what experimentation teams can do about it



Kyle Newsam

Experimentation Lead, Mando Group

Six reasons experiments come back inconclusive

Intro	The uncomfortable reality of experimentation	03
01	You're measuring the wrong metric	04
02	You're testing too little traffic	05
03	You're including people who never actually experience the change	06
04	Your hypothesis isn't strong enough	08
05	You're trying to answer too many questions with one experiment	09
06	Your foundations shape every experiment	10
→	Bringing it together: Design for conclusive learning	11
→	Pre-launch checklist	12
→	Conclusion	13



The uncomfortable reality of experimentation

A large proportion of experimentation programmes produce inconclusive results. Optimizely's analysis of 127,000 experiments found that only 12% produce a statistically significant improvement on the primary metric, and it describes a conclusive rate of 35 to 40% as healthy.

That can create a problem for experimentation leaders.

If the majority of experiments are coming back as "no significant difference", it becomes harder to demonstrate the value of experimentation to senior stakeholders. It can also create the perception that experimentation itself isn't working.

But an inconclusive experiment doesn't necessarily mean the underlying idea was bad.

Sometimes the experiment was simply not designed in a way that gave it a strong opportunity to detect an effect.

12%

of experiments produce a statistically significant improvement on the primary metric.

Source: Optimizely, Top 10 takeaways from running 127,000 experiments



This guide explores some of the less obvious reasons experiments become inconclusive, and how teams can improve their experiment design to increase the proportion of experiments that produce useful, actionable outcomes.



01 You're measuring the wrong metric

One of the most common mistakes is selecting a primary metric based on what the business ultimately cares about rather than what the experiment is actually trying to change.

For example, an e-commerce business might use revenue as the primary metric for almost every experiment.

That sounds logical. Revenue matters to the business.

But revenue is also influenced by a huge number of variables, and the further away a metric is from the behaviour being directly influenced by the experiment, the harder it can be to detect an effect.

EXAMPLE

Imagine you're testing a new product listing page. The immediate behavioural objective might be to encourage more users to explore individual products. The behavioural sequence could look something like:



If the experiment changes the listing page, the most immediate behavioural impact is likely to be movement from the listing page to the product detail page.

Trying to measure the experiment primarily through revenue introduces multiple additional stages where other factors can influence the outcome.

Instead, the primary metric could be product-detail-page click-through rate, while downstream metrics such as purchase rate and revenue remain important secondary metrics.

THE PRINCIPLE

Measure the behaviour you're trying to influence, not just the business outcome you ultimately care about.

This can give the experiment a better chance of detecting whether the change actually worked.



02 You're testing too little traffic

Another common issue is unnecessarily restricting an experiment to a small amount of traffic.

This often happens when teams test an individual page even though the experience they're changing exists across a much larger number of pages.

EXAMPLE

An e-commerce website might have thousands of product detail pages using essentially the same template.

If a team tests a change on one individual product page, the experiment may struggle to collect enough observations to determine whether the change has an effect.

But if the hypothesis concerns a common component of the product-detail-page template, the experiment could potentially be applied across the broader set of eligible product pages.

Instead of:

One product



Limited traffic



Inconclusive result

You can potentially test:

Product-detail-page template



Thousands of products



Substantially more traffic

Once the experiment has sufficient statistical power, teams can then investigate whether the effect varies across different product categories, customer groups or other meaningful segments.

THE PRINCIPLE

If the experience is shared, don't unnecessarily restrict the experiment to one instance of it.



03 You're including people who never actually experience the change

This is a subtler problem: experiment exposure.

A user can technically be included in an experiment without ever seeing the variation being tested.

Imagine you're testing a new component halfway down a webpage. A user lands on the page, gets assigned to the experiment, but never scrolls far enough to see the component. That user has technically been exposed to the experiment from the platform's perspective. But behaviourally, they were never exposed to the change.

Including large numbers of these users can dilute the measured effect, particularly when the experiment is testing an element that sits below the fold.

✓ The solution: trigger on meaningful exposure

Rather than assigning someone to the experiment as soon as they land on the page, teams can potentially configure a custom trigger so that the user enters the experiment when the relevant element actually becomes visible.

For example:



This can provide a much closer relationship between being included in the experiment and actually experiencing the treatment.

Continued overleaf →



This is typically a more advanced experimentation configuration rather than something teams necessarily need to implement for every test.

The first question should still be:

② Why are we testing something that most users don't see?

For many experiments, particularly those intended to influence initial behaviour, prioritising changes within the first viewport can make more sense.

However, when there is a legitimate reason to test a below-the-fold component, exposure-based triggering can become an important part of experiment design.



Important caveat

This shouldn't become an excuse to manipulate the experiment population until the results look better.

Teams should define what constitutes meaningful exposure before analysing the results and ensure the triggering logic is consistent and defensible.

THE PRINCIPLE

Being assigned to an experiment isn't the same as experiencing the experiment.



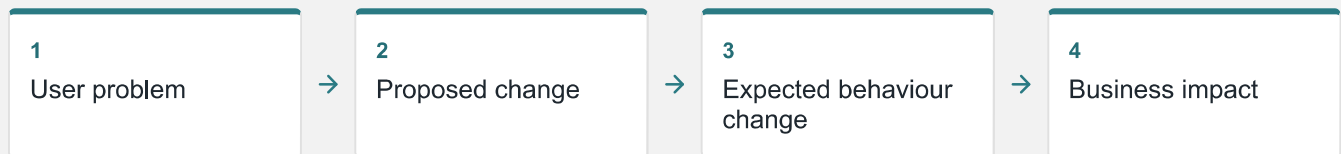
04 Your hypothesis isn't strong enough

Not every hypothesis deserves an experiment.

Sometimes the underlying change simply isn't substantial enough to meaningfully alter behaviour.

For example, changing a minor visual detail that has little connection to the user's decision-making process may technically produce a measurable difference, but the expected effect could be so small that detecting it reliably requires an impractical amount of traffic.

A strong hypothesis should establish a clear relationship between:



For example:

✓ STRONG HYPOTHESIS

"Users may hesitate to proceed because they cannot easily understand the delivery timeframe. Making delivery information more prominent should increase progression from product detail to checkout."

This provides a much stronger basis for experimentation than:

✗ WEAK HYPOTHESIS

"Making the delivery information bigger will improve conversion."

THE PRINCIPLE

The stronger the causal connection between the change and the expected behaviour, the more meaningful the experiment is likely to be.



05 You're trying to answer too many questions with one experiment



Experiments can also become inconclusive when the intervention is too complicated.

If an experiment changes the navigation, page structure, copy, imagery and CTA simultaneously, the team may eventually see a difference, but have little understanding of what caused it.

Conversely, if several changes interact with one another, the overall effect may become unpredictable.

Teams should therefore think carefully about the scope of the intervention.

The experiment should be large enough to create a meaningful behavioural difference, but structured enough that the hypothesis remains understandable.

THE PRINCIPLE

An experiment should answer a question, not create five new ones.



06 Your foundations shape every experiment

The same foundations that give experiments a strong chance of a clear result also decide how much value AI can add to your experimentation practice.

AI doesn't magically fix weak experimentation infrastructure.

The organisations getting the most value from AI are likely to have strong foundations: structured data, good analytics, accessible research, established design systems, clean repositories, consistent documentation and clear governance.

Structured data

Good analytics

Accessible research

Established design systems

Clean repositories

Consistent documentation

Clear governance

AI can amplify those foundations. It can also amplify their absence.

If your data is fragmented and your experimentation repository is a graveyard of forgotten PowerPoint decks, adding AI doesn't necessarily solve the problem.



**You have simply
given the chaos a
chatbot.**



BRINGING IT TOGETHER

Design for conclusive learning

The goal shouldn't be to eliminate inconclusive experiments.

Experimentation inherently involves uncertainty. Some hypotheses will be wrong. Some effects will be too small to detect. Some experiments will simply fail to produce a meaningful difference.

The goal is to reduce avoidable inconclusiveness.

Before launching an experiment, teams should therefore ask the questions in the checklist below.

[Go to the pre-launch checklist →](#)



Before you launch

Hypothesis

- ☐ Is there a meaningful user problem or opportunity?
- ☐ Is there a credible mechanism through which the change could influence behaviour?

Metric

- ☐ Does the primary metric represent the behaviour the experiment is actually trying to change?
- ☐ Are downstream business metrics being used appropriately as secondary measures?

Traffic

- ☐ Is the experiment receiving enough traffic?
- ☐ Could the same hypothesis be tested across a broader reusable template or experience?

Exposure

- ☐ Are users actually experiencing the change?
- ☐ If the element is below the fold, should experiment assignment be triggered by meaningful exposure?

Experiment design

- ☐ Is the intervention appropriately scoped?
- ☐ Are we introducing unnecessary complexity or multiple unrelated changes?

Foundations

- ☐ Can we find what we learned from similar experiments in our repository?
- ☐ Do we trust the analytics behind the primary metric?

An inconclusive result isn't necessarily a failed experiment.

Sometimes it is telling you that there wasn't a meaningful effect.

But sometimes it is telling you something about the way the experiment was designed.

The difference matters.

The most mature experimentation programmes don't simply focus on increasing the number of experiments they run. They focus on increasing the quality of the questions they ask, the populations they expose, the metrics they measure and the conditions under which they run those experiments.

Because the objective isn't to manufacture more statistically significant results.

It's to create experiments that give you the best possible opportunity to learn something useful.



TALK TO US

At Mando Group, our experimentation specialists work alongside your team to design experiments that reach a clear answer and build the foundations that make AI useful.

Talk to us about Optimise →

mandogroup.com/get-in-touch

Source

Optimizely, Top 10 takeaways from running 127,000 experiments



MANDO3ROUP