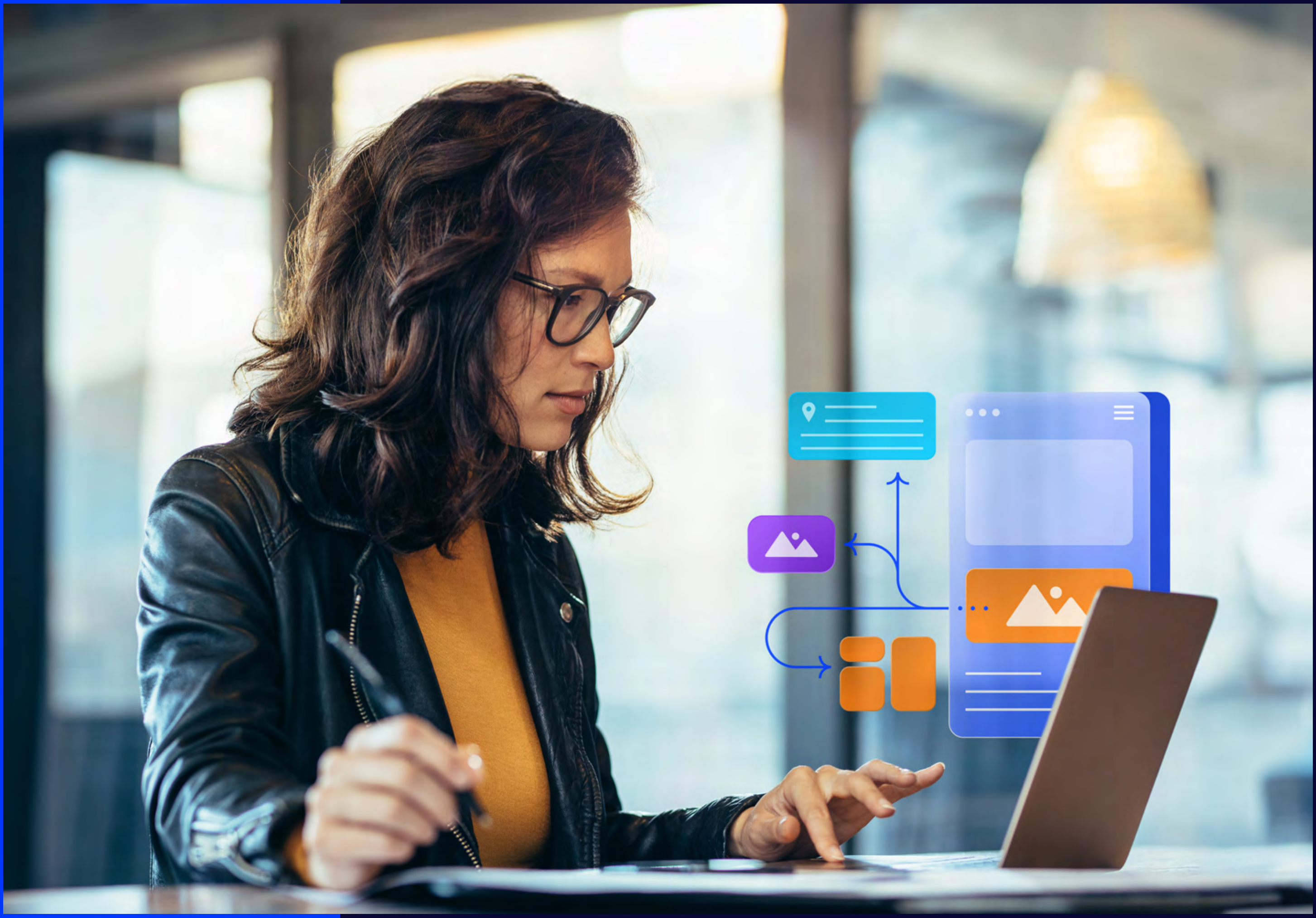


EBOOK

Evolution of experimentation

Elizabeth Gabster, Senior Director, Strategy & Value Advisory
Eric Lang, Senior Consultant, Strategy & Value Advisory
Kory Manley, Senior Director, Strategy & Value Advisory
Hazjier Pourkhalkhali, Global VP, Strategy & Value Advisory
Emma Shillam, Lead Consultant, Strategy & Value Advisory
Mark Wakelin, Lead Consultant, Strategy & Value Advisory

–

Optimizely

Introduction

Business experimentation is based on one central tenet: that applying a scientific lens to business processes can help us make better decisions than our past ways of working.

Yet this same scientific rigor is rarely applied to the craft of experimentation itself. Although our profession is now over a decade old, there is still limited public information available for practitioners to know what will truly make them successful or for executives to know how to drive business outcomes.

We believe that as the world's largest digital laboratory, we have a responsibility to our industry and clients to share our data and best knowledge for good. We hope that the insights that we share here can help the next generation of practitioners and business leaders make advances faster and invent better ways of working.

In creating this report, we reviewed over five years of experiment data on Optimizely and included scientific research from academics at Harvard Business School that we want to share with a wider audience.

> “All life is an experiment. The more experiments, the better.”
>
> **Ralph Waldo Emerson**, 1844

Contents

01 About our research

Our work is based on four sources of data

This analysis of the experimentation landscape includes not only benchmarks of our experiment data set, but also valuable insights from other sources to supplement the findings.

1 **2023 Optimizely experimentation benchmark**

Analysis of over 127,000 true experiments conducted on Optimizely Web Experimentation and Optimizely Feature Experimentation between 2018 and 2023.

2 **Excerpted Optimizely analyses**

Select Optimizely analyses conducted over the years that are separate from the above benchmark, but are included here as there are valuable insights for the broader industry.

3 **Customer interviews, case studies, and surveys**

Interviews of key accounts, selected case studies, and surveys conducted on Optimizely customers in 2023.

4 **Academic research**

Excerpts from scientific research on experimentation, both conducted on Optimizely's data as well as outside in.

Three key definitions for our research

As part of our research, we use the following terms frequently:

True experiments

A true experiment is a correctly set up test in a production environment with sufficient traffic and real variations. That requires a control variation with ≥1,000 visitors, one or more treatment variations with ≥1,000 visitors, and no signs of the treatments being A/A, meaning Web variations that include code or feature experimentation variations that do not have A/A naming conventions.

Winning experiments

Scientifically, experiments do not win, they merely disprove the null hypothesis. However, in our industry, practitioners often speak of winning experiments and we continue those terms to be more easily understood. Winning experiments are true experiments, with the metric in question moving in the "winning direction" (98% of the time this is uplift) and statistical significance (≥90%).

Expected impact

Expected impact is the expected value of an experiment, meaning: how likely is the experiment to win × what is the winning uplift?

Example: 10% win rate × 10% uplift = 1% expected impact

We use this to forecast future returns for programs and to avoid over-indexing on high win rates with low uplifts.

↓

Detailed definitions and exact examples are provided on pages 55-59

Key Takeaways

The state of experimentation

1. Around 12% of experiments win on the primary metric. 88% of tests do not win.
2. The median company runs 34 experiments per year. The top 3% of companies run over 500 experiments per year.
3. The number of companies testing, their experimentation velocity, and the share of feature experimentation have consistently grown since 2018.
4. Companies ramp up testing quickly from launch and grow velocity by 20% year over year on average.
5. Most experiment uplifts decay to 80% of their first month value after a year, except for revenue, which retains 91% after a year.

Great experiments

6. The performance of teams is stable over a three-year timeframe. Improving in performance requires continually changing the system by which you research, ideate, and develop experiments.
7. The highest uplift experiments around the world have two characteristics in common: they test a higher number of variations and implement more complex changes.
8. Less than 10% of experiments test 4 or more variants, yet those experiments are more than twice as impactful compared to A/B.
9. Only a third of experiments make more than 1 change, yet they show much better returns.
10. Digital commerce overwhelmingly prioritizes revenue. Huge early funnel optimization opportunities like search and add-to-cart are underexplored.
11. Personalized experiments drive 41% more expected impact.

Great cultures of experimentation

12. Great data makes the difference. Companies who use advanced analytics are far more successful at experimentation.
13. Companies with an integrated CDP appear to be much more successful with experimentation.
14. Senior leaders tend to be involved with more winning experiments yet have smaller uplifts. Great leaders should encourage teams to take risks and explore alternatives.
15. Large programs appear evenly split between centralized and decentralized teams, with limited performance difference observed.

02
The state of experimentation

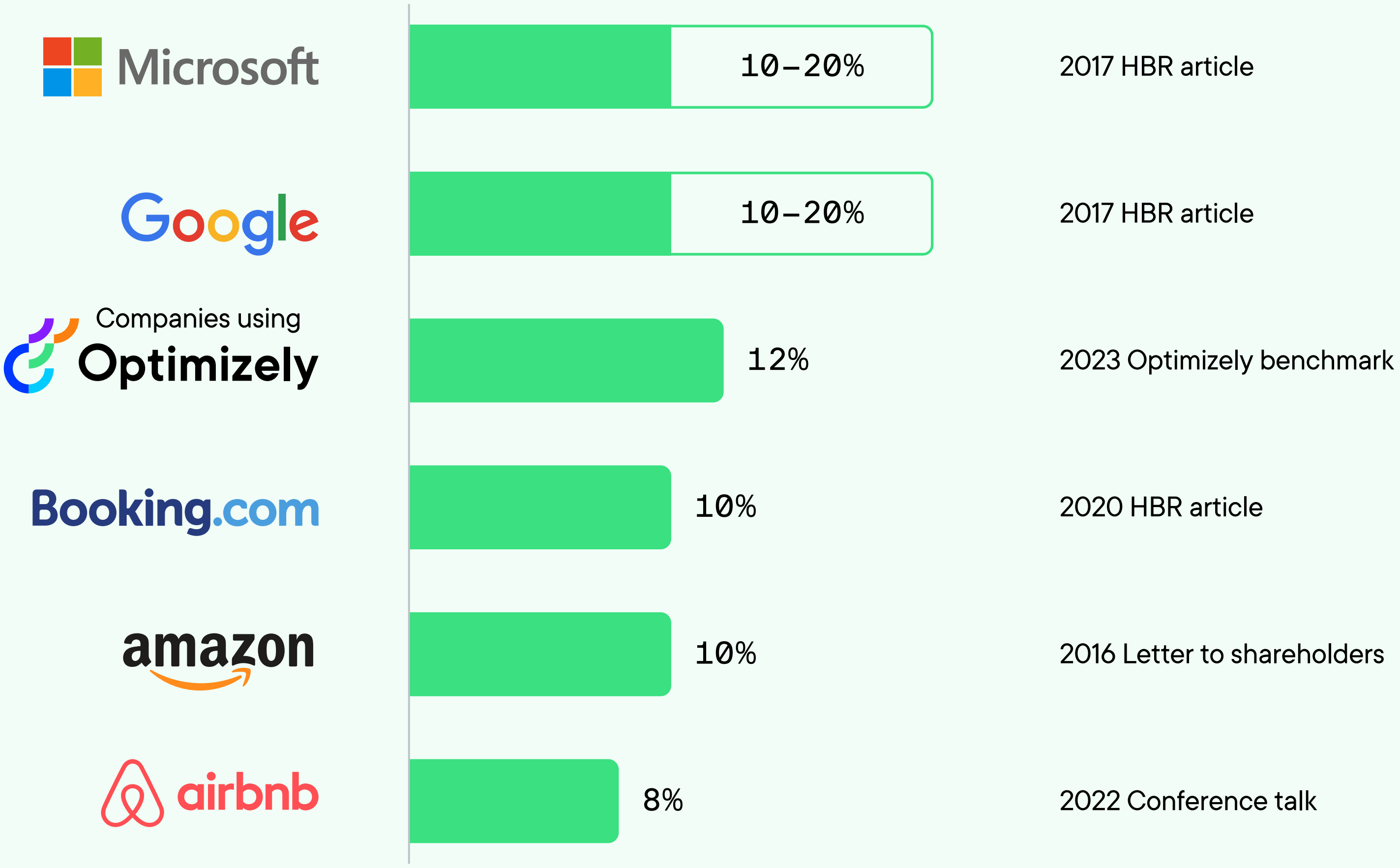
Changing your website is easy, changing user behavior is hard

The data is clear: for every 8 to 10 updates, feature releases, or design changes that companies launch, only one changes user behavior for the better. Growth is not a function of moving quickly, but distinguishing what works from what doesn't.

> "Given a 10 percent chance of a 100 times payoff, you should take that bet every time. But you're still going to be wrong nine times out of ten."
>
> **Jeff Bezos**, 2016
> Letter to Shareholders

Around 1 in 8 experiments will win on the primary metric

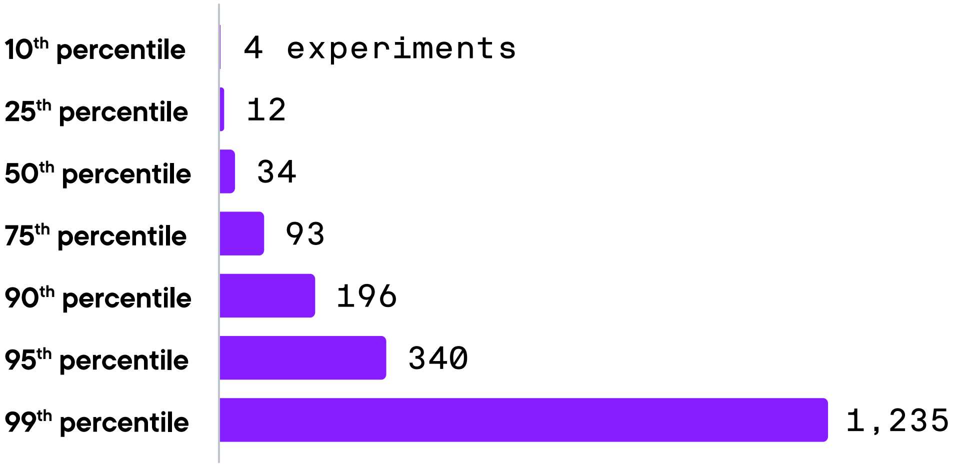
Experiments achieving a statistically significant improvement, self-reported



To be in the top 10% of experiment velocity, companies need to run around 200 tests annually

Experiment velocity by company

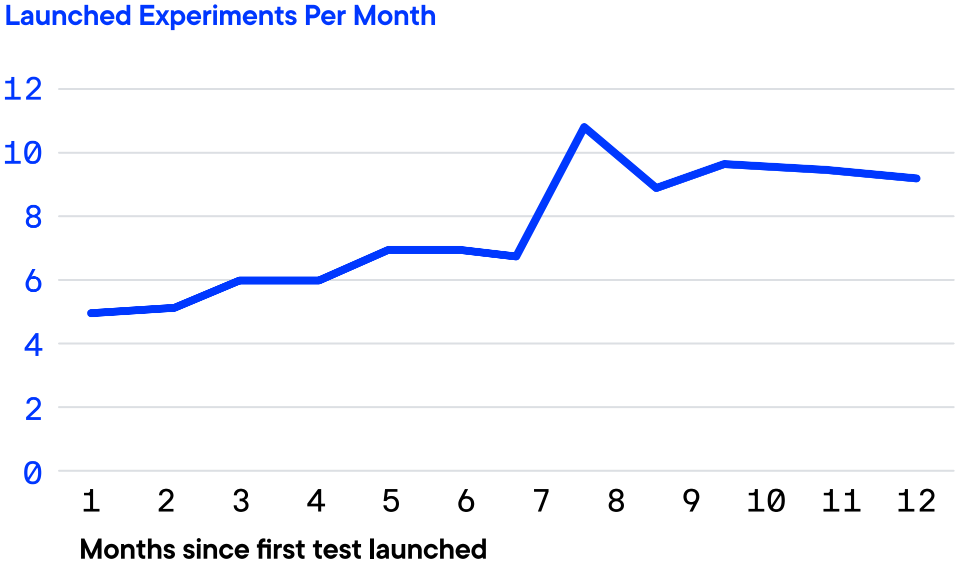
Experiments created in 2022, over 900 companies



- The median company runs around 3 experiments per month.
- Reaching the top 10% of velocity requires scaling to 16+ tests per month.
- Only 3% of companies are in the elusive 500 tests club.
- The top 1% run over 1,000 tests per year.

Experiments created per month from first test launch on Optimizely

116 companies creating their first experiment 1st December to 1st June 2022



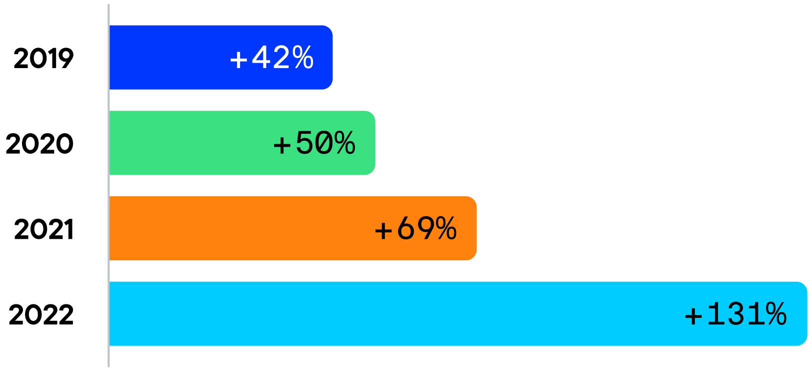
- Companies gradually ramp up their velocity in the first year.
- Despite a spike observed around 3 quarters in, the trendline is relatively consistent.
- Longer-term analyses show a roughly 20% increase in velocity from the first to the second year.

Digital maturity is growing over time

The adoption of experimentation has grown substantially in the past five years. We see companies increasing in their velocity and the share of companies testing rising. We believe this shows experimentation maturing from a niche business practice for early adopters to a standard expectation across more and more companies.

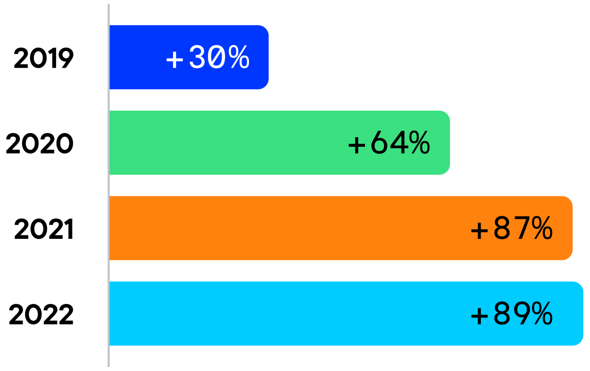
Number of experiments started in Optimizely per year

Increase in number of experiments compared to 2018



Number of companies starting experiments in Optimizely per year

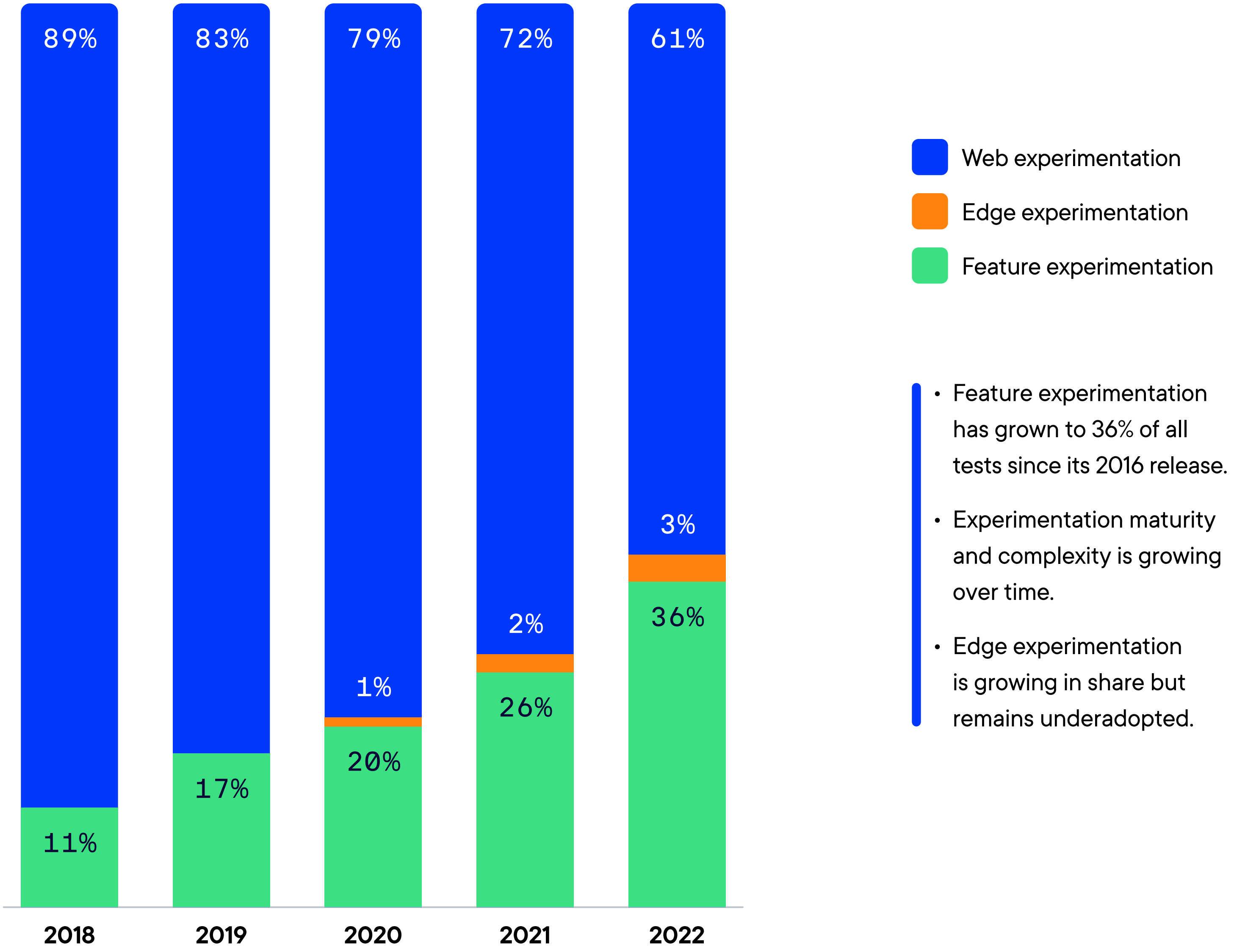
Increase in number of companies compared to 2018



Companies are increasingly moving from client-side testing to more mature experimentation frameworks

Share of experiments by channel over time

n = 120k true experiments run 2018-2022 across 1.1K companies

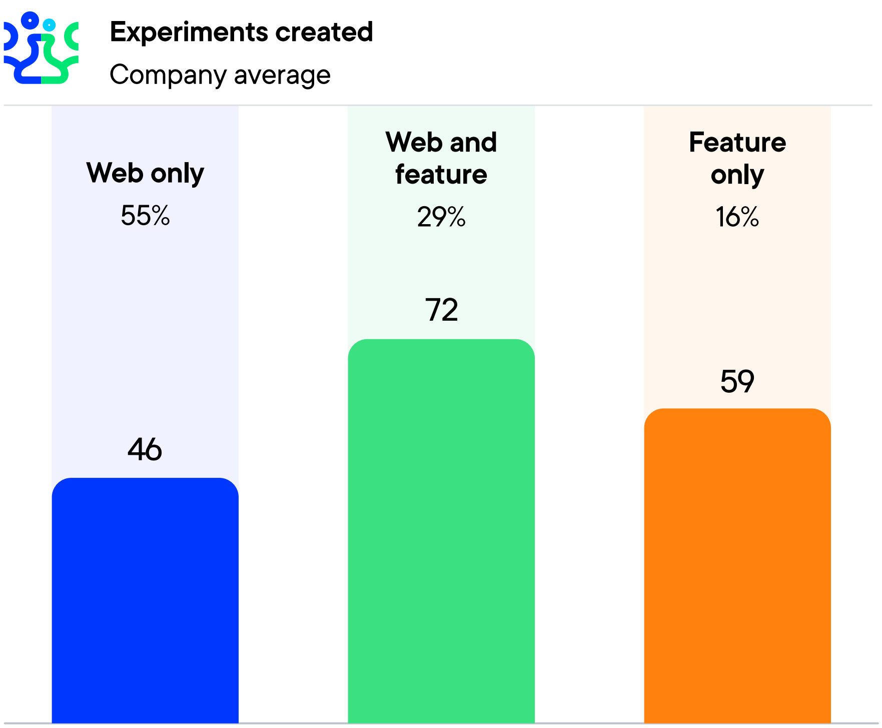


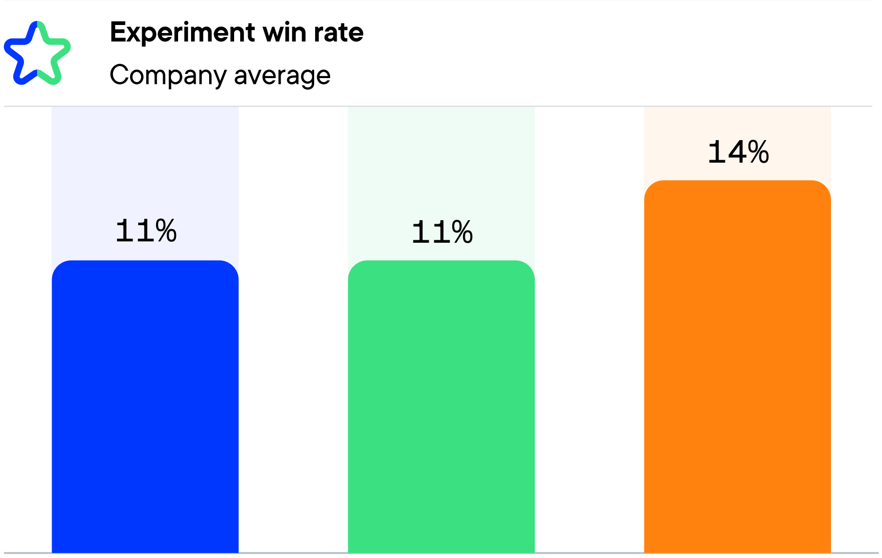
- Feature experimentation has grown to 36% of all tests since its 2016 release.
- Experimentation maturity and complexity is growing over time.
- Edge experimentation is growing in share but remains underadopted.

Feature Experimentation outperforms Web

Company experiment performance, by client/server side

Experiments in 2022 across companies with 12+ tests in 2022





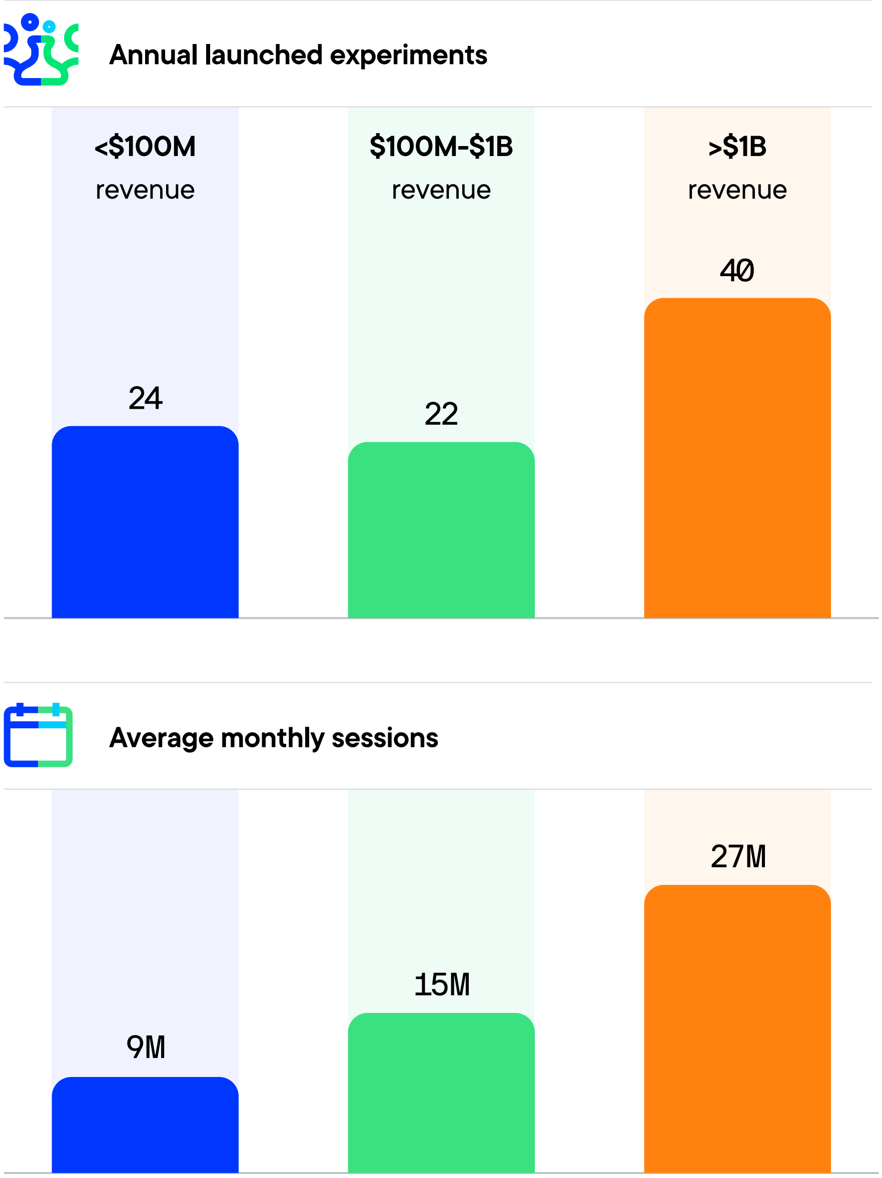
Flexibility drives velocity: Companies with deployment flexibility face fewer obstacles and can scale under more scenarios.

Technical maturity drives success: Organizations that adopt more advanced technologies show better experiment performance.

There appear to be some competitive advantages for companies with over $1B of revenue per year

Company revenues, traffic and program metrics

Experiments in 2022, Similarweb traffic data for 2022



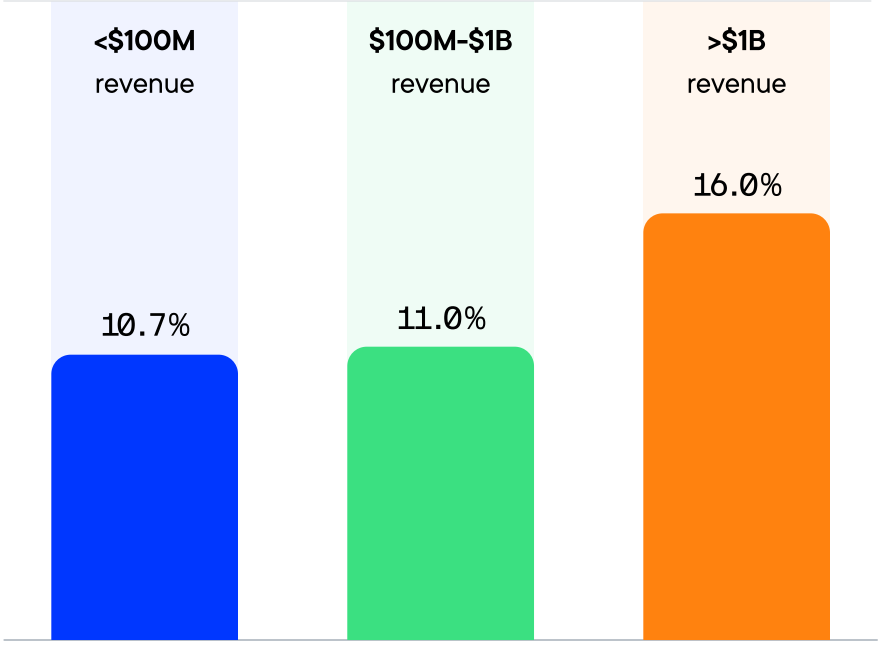
Annual launched experiments

Average monthly sessions

Revenue matters: Larger companies have more resources and are more likely to run high velocity programs.

Data is fuel: Large companies generally have higher visitor counts and hence they can measure test results more accurately.

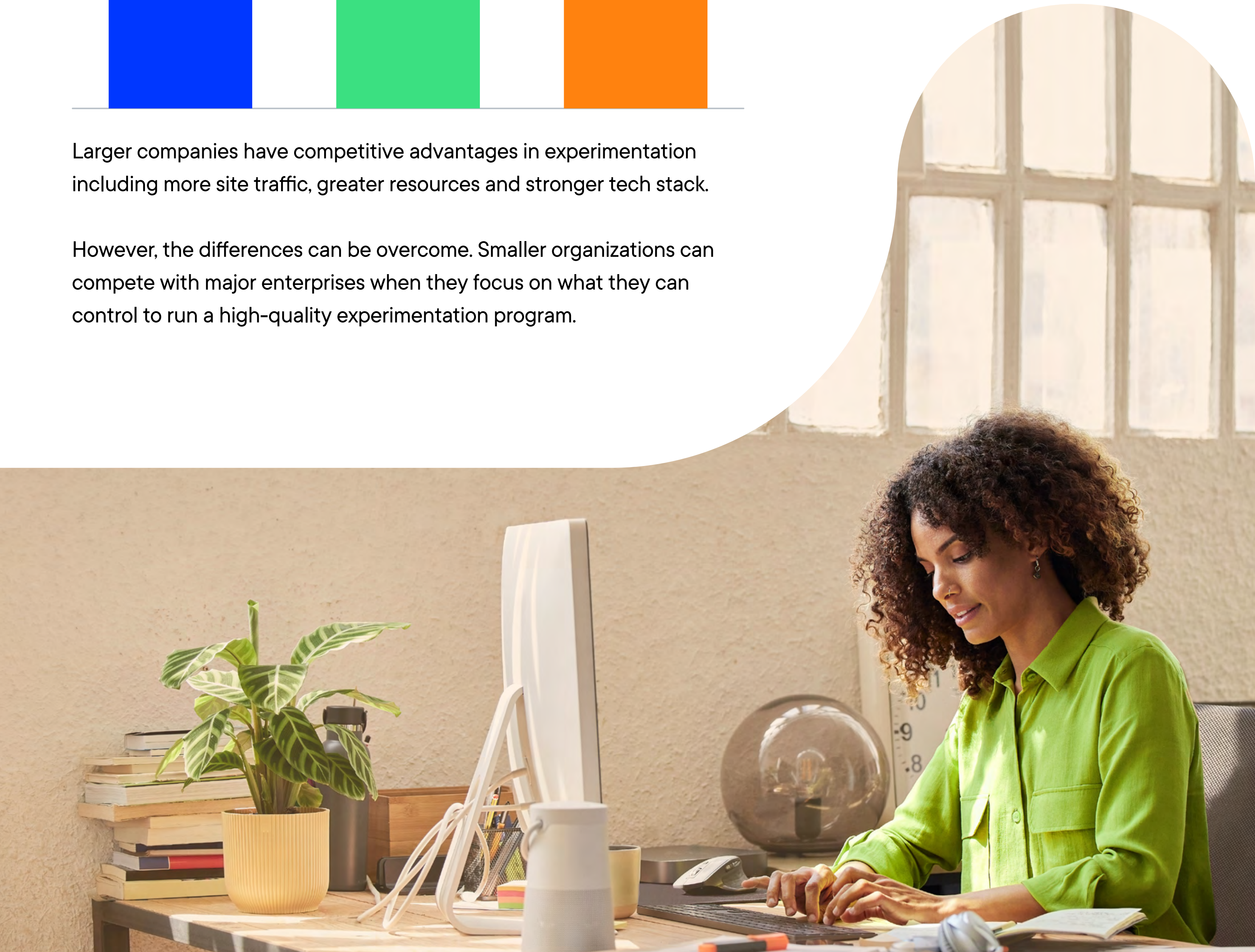
Experiment win rates



Larger companies have competitive advantages in experimentation including more site traffic, greater resources and stronger tech stack.

However, the differences can be overcome. Smaller organizations can compete with major enterprises when they focus on what they can control to run a high-quality experimentation program.

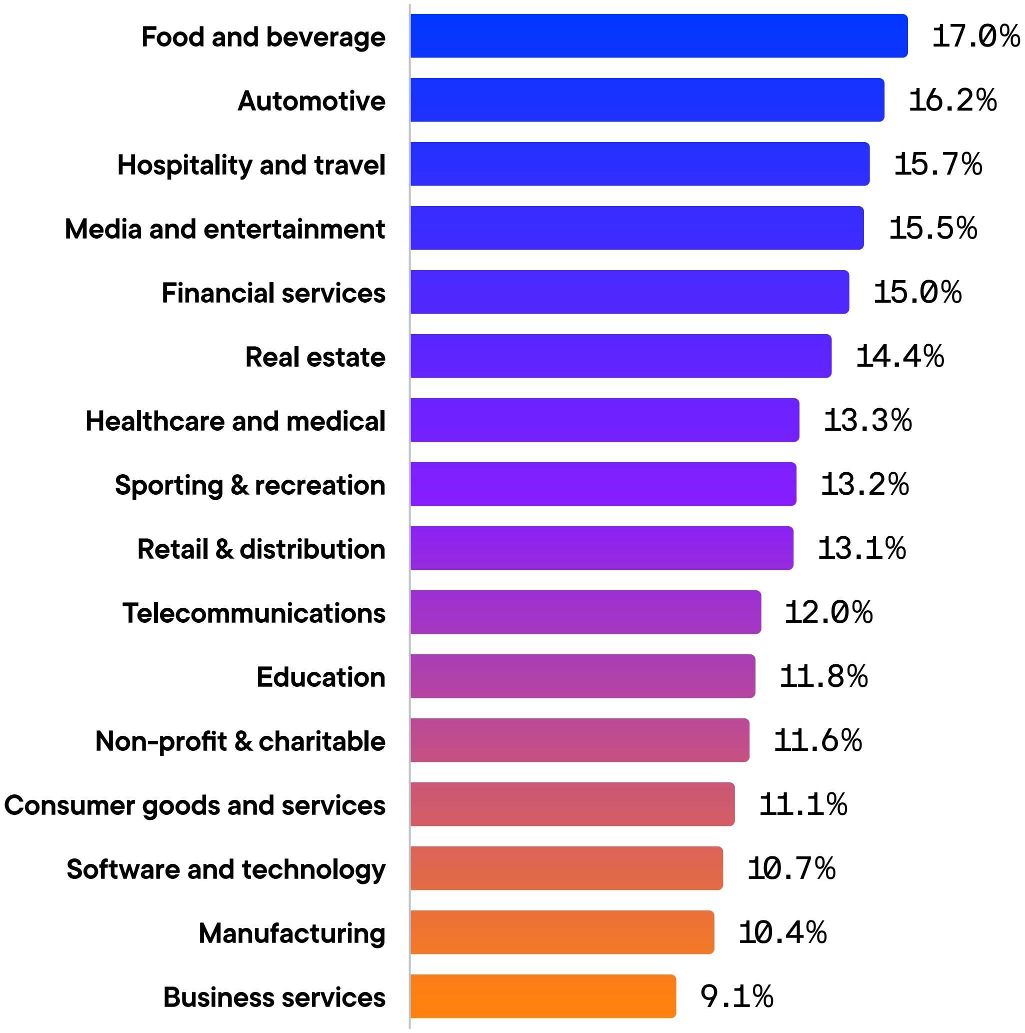
Win rates differ:
At higher revenues, companies appear to enjoy some degree of competitive advantage in their experiment win rates.



There are modest differences in win rates between industries, and may be explained by experimentation maturity and metric selection

Experiment win rates by industry

Win rates on primary metric for true experiments, industries with >15 companies



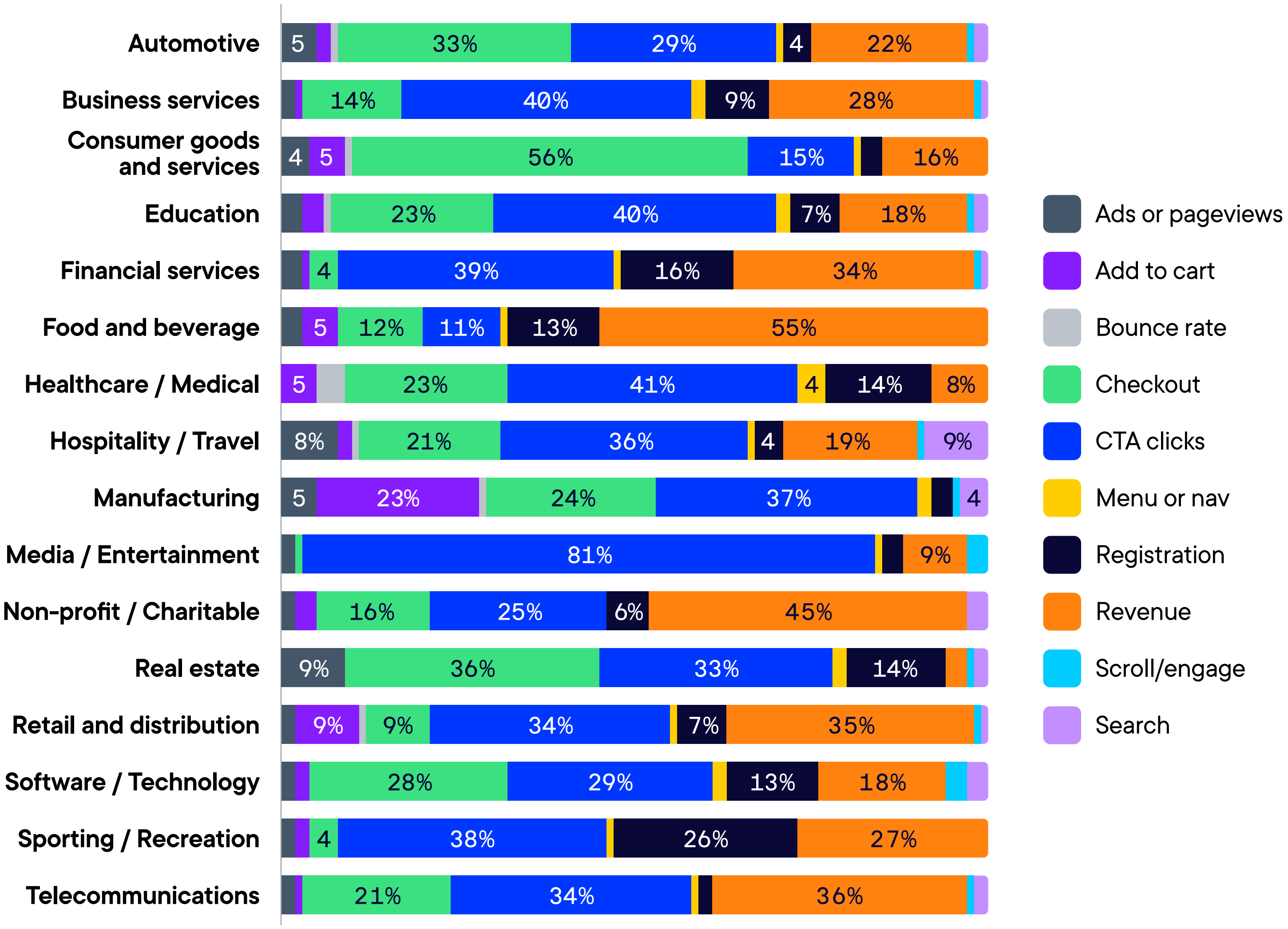
Industry differences are not destiny: High and low performers exist in all industries. And all companies have potential to improve their performance.

Averages only: These results only show averages of industries, and do not adjust for company characteristics or maturity. Future research will investigate where industry differences come from.

Primary metrics vary by industry, due to differences in goals, priorities, and tracking capabilities

Primary metric share by industry

True experiments, n = 1.1k companies, n = 127k experiments, industries with >15 companies



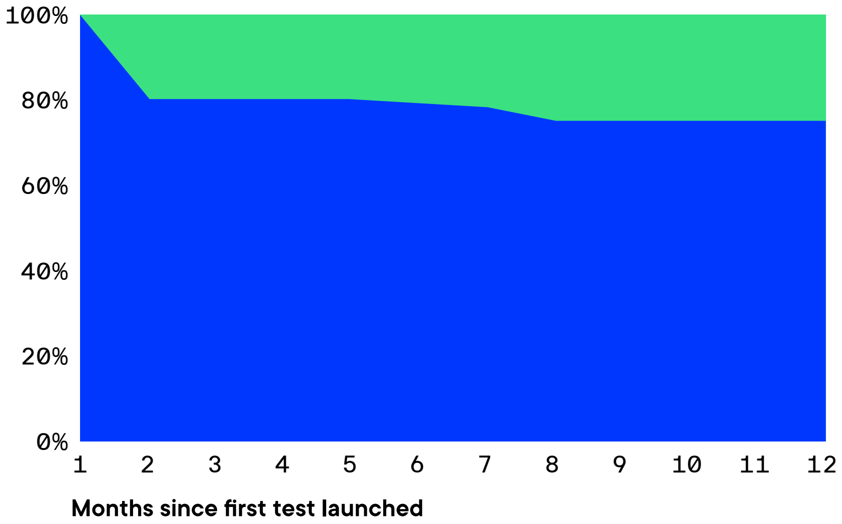
Companies need to factor in regression to the mean when estimating future value from experiments

Proportion of the first month's uplift that is retained every month
Uplifts on winning, true experiments run for >12 months,
Optimizely analysis, 2019

🟦 Uplift retained over a year

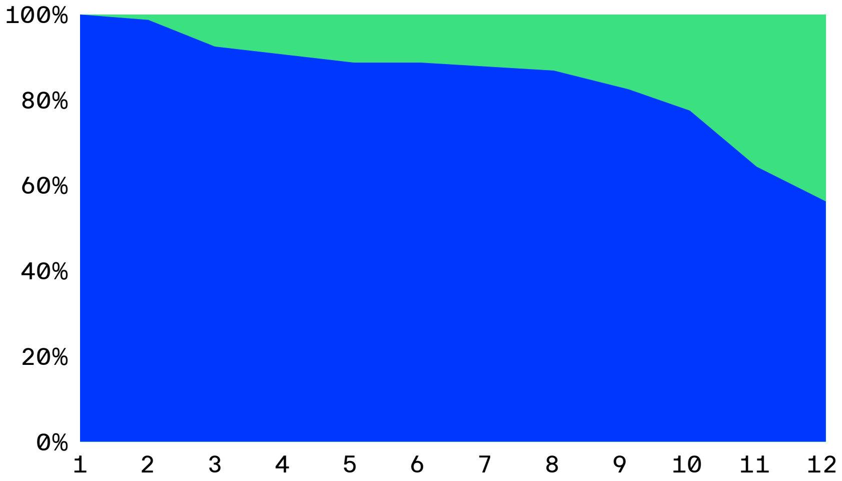
🟩 Uplift decay over a year

Uplift on click goals



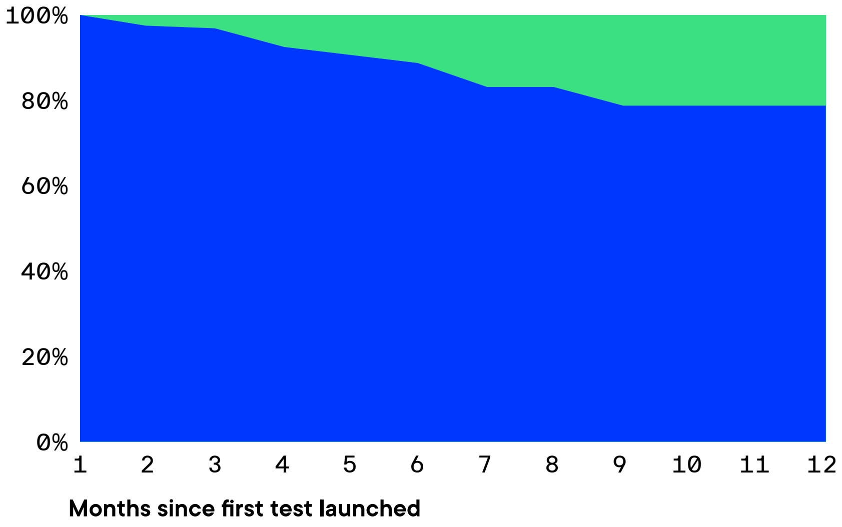
Click goals often decay early and stabilize quickly, retaining **79%** of the uplift after a year.

Uplift on custom metrics



Custom metrics decay continually and retain only **57%** of uplift after a year.

Uplift on pageviews

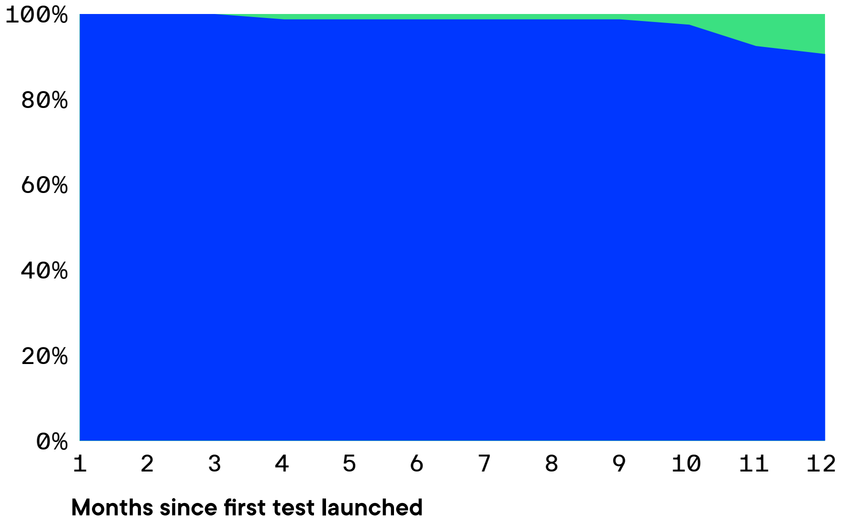


🟦 Uplift retained over a year

🟩 Uplift decay over a year

Pageviews decline gradually and retain **75%** of their uplift after a year.

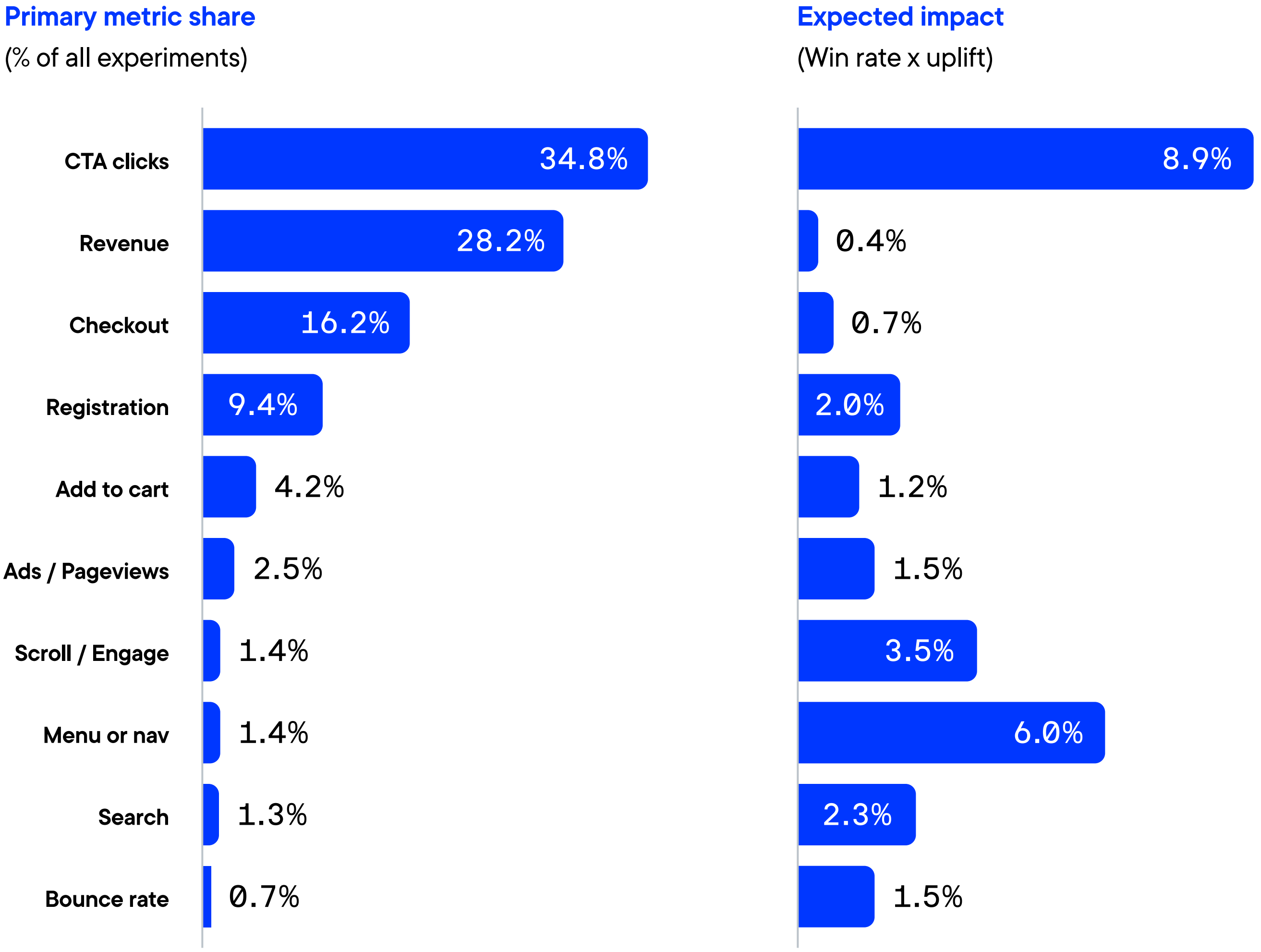
Uplift on revenue metrics



Revenue metrics show the least decay and retain **91%** of the uplift after a year.

Over 90% of all experiments target the top 5 metrics

Primary metric share and expected impact across all experiments
True experiments, n = 1.1k companies, n = 40k experiments



The most common experiment around the world is optimizing a call-to-action, which also carries the highest expected impact of any metric.

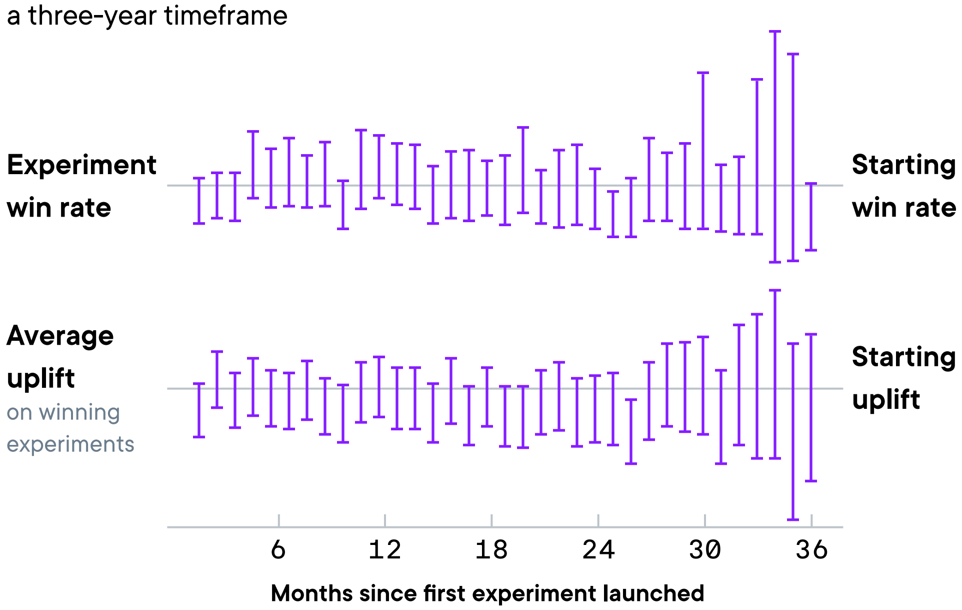
8.9% expected impact = 23.4% win rate × 38.2% average winning uplift

Over 92% of all experiments target the Top 5 metrics. As a result, serious optimization opportunities around improving menu / navigation and site search are underprioritized by most programs.

03 Characteristics of great experiments

Unless you change your patterns, however good you are today is likely how good you are in 3 years time

Scientific research with Stanford:
The success rates of experimentation teams remain stable over a three-year timeframe



Source:
Preliminary analyses (Stanford's Graduate School of Business and Optimizely, 2021) n = 76,167 experiments from 1,589 companies.

Testing teams remain relatively consistent in win rates and uplifts over a three-year timeframe. This suggests that teams do not face diminishing returns over time; they continue to be as productive years later.

However, it also suggests that teams can be stuck in comfort zones. Testing more often and accumulating experience will not improve performance by itself.

To advance to a higher level of productivity, you must apply your knowledge by changing the system through which you experiment. That requires improving the depth of research, tracking more user behavior, gaining access to development resources, being more willing to take risk, and finding freedom to pursue novel ideas.

Based on our research, teams can structurally improve their performance by focusing on five elements:

The world's highest performing experiments follow a very different pattern than the usual norm of "A/B Testing." Contrary to industry advice that good experiments should be minute in scope and limited to two variants, the data shows that the highest performing experiments have very different characteristics.

Learnings from the best experiments run on Optimizely

Explore more options:
High variant tests outperform

Companies can find up to 5x more wins by focusing on micro-conversions.

Personalize to your users:
Targeting increases success

Experiments that make major changes to the user experience are more likely to win and with higher upfits.

Think bigger:
Complex experiments drive more impact

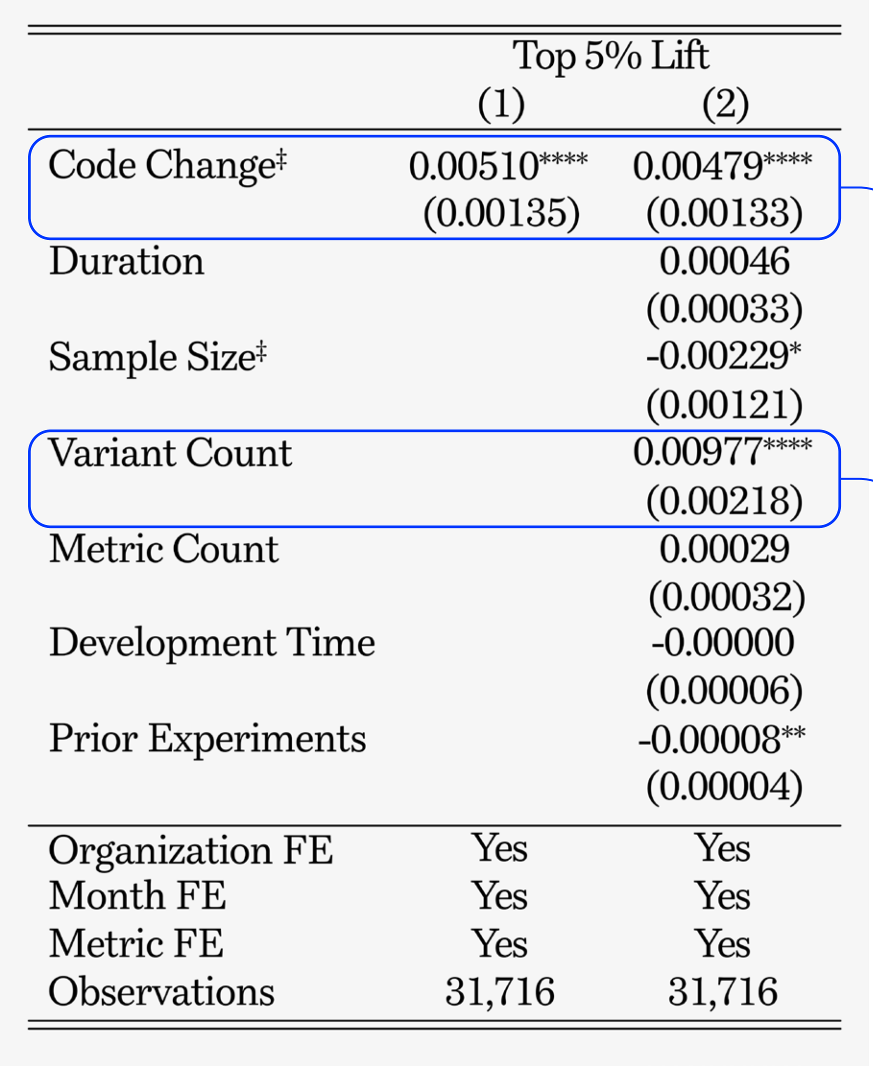
Experiments that test multiple treatments are 3x more successful than A/B tests.

Set the right goals:
Choose the right metrics for success

Experiments leveraging bandit algorithms are more successful.

The experiments with the highest uplifts make substantial code changes and test many variations

Scientific Study with Harvard Business School: The biggest breakthrough experiments have a high degree of code change and test many variations



	Top 5% Lift	
	(1)	(2)
Code Change‡	0.00510****	0.00479****
	(0.00135)	(0.00133)
Duration		0.00046
		(0.00033)
Sample Size‡		-0.00229*
		(0.00121)
Variant Count		0.00977****
		(0.00218)
Metric Count		0.00029
		(0.00032)
Development Time		-0.00000
		(0.00006)
Prior Experiments		-0.00008**
		(0.00004)
Organization FE	Yes	Yes
Month FE	Yes	Yes
Metric FE	Yes	Yes
Observations	31,716	31,716

Experiments in the top 5% of uplifts have two characteristics in common:

1 They make larger code changes with more effect on the user experience (>99.9% significance).

2 They test a higher number of variations simultaneously (>99.9% significance).

This suggests that great experiments need to try large leaps in the user experience balanced with an openness to multiple paths.

Source: Ghosh, Sourobh. 2021. Experimental Approaches to Strategy and Innovation. Doctoral dissertation, Harvard Business School.

“The standard experiment run around the world is an incremental A versus B test. While these tests are easy to run, they rarely associate with performance breakthroughs.

Our data shows that the largest breakthroughs come from tests that follow a very different model. Tests which are designed to test complex, interdependent changes–but within a single variant and across multiple variants–are more likely to be among the top 5% performing experiments in our sample.

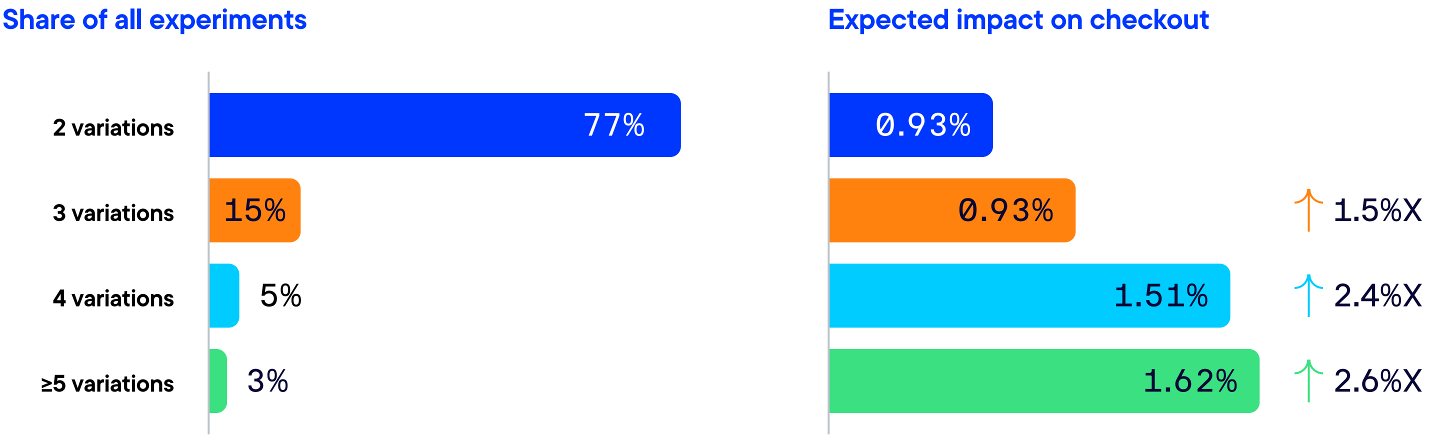
Rather than shying away from complexity, firms can potentially harness it to deliver high performance in testing. The key is to pair complex tests with a theory for how the multiple elements work together to deliver returns. Theory and testing together can help firms unlock breakthrough performance.”

Dr. Sourobh Ghosh, Economist at Amazon / Audible
Ph.D in Business Administration from Harvard Business School

Companies overwhelmingly test A versus B, when the highest value is for multiple variants

Experiments by variants versus expected impact on checkout

n = 127K tests, variations includes original (2 Variations = Original + 1 Treatment)



Share of all experiments

Expected impact on checkout

How do we change as we test more variants?

- **Teams take more risks as safe options are covered**
  All teams care about having winning tests. With a single variant, teams often play it safe. But when teams test 4+ variations, the safe options are covered and increasingly risky and novel ideas can be tested without worry.

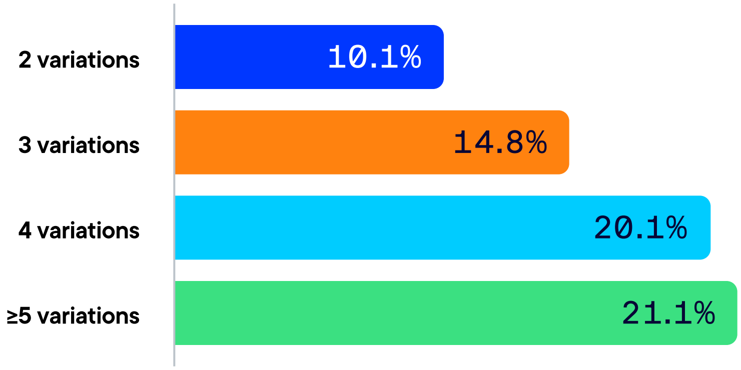
- **There is greater ownership and likelihood to contribute**
  Teams that test A versus B often choose their B variant through hierarchy or design by committee. When teams test multiple variants, people see a chance to participate and have their ideas tested. Their job is no longer to measure a change, but to increase the likelihood of a change succeeding.

- **Programs become more agile and open-minded**
  In a waterfall model, teams can test only a single variant at a time because there's only one path to follow. If teams are agile, they test many variations simultaneously and change their direction based on the results.

Across primary metrics, higher variants outperform A/B

Experiment win rates by number of variations (including baseline)

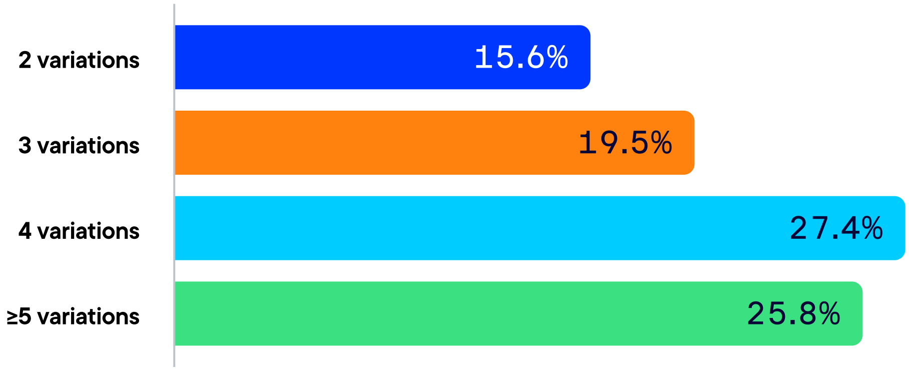
Win rate on the primary metric for true experiments, n = 127k tests



- Over 77% of all experiments test only 2 variations (original + treatment).
- Tests with 4 or more variations are twice as likely to win on average.

Average uplift for winning experiments by variations (including baseline)

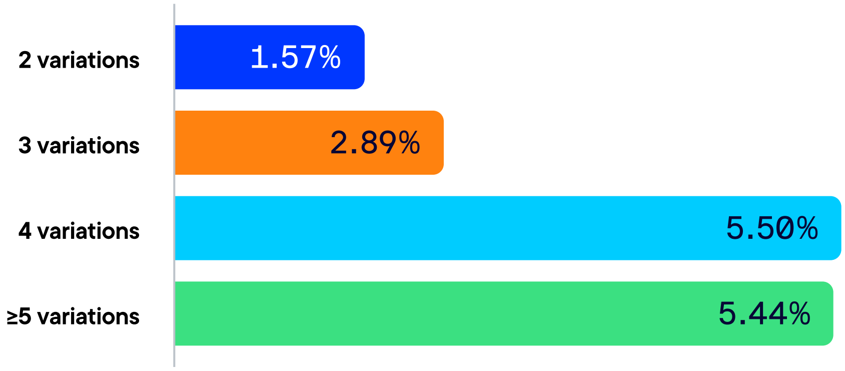
Average winning uplift on the primary metric for true experiments, n = 127k tests



- When higher variant tests win, they result in much larger uplifts.
- While testing more variants requires more traffic, these higher lifts mitigate the traffic needed because larger uplifts can be detected with less traffic.

Expected impact by number of variations (including baseline)

Expected impact on the primary metric for true experiments, n = 127k tests



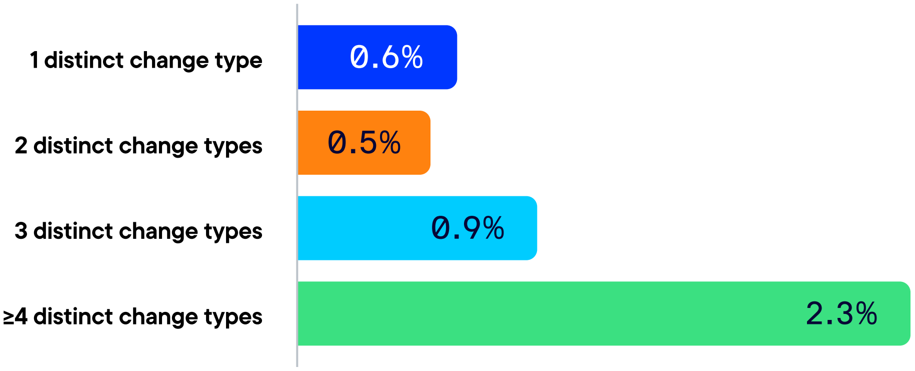
- Because higher variant experiments have higher win rates and uplifts, experiments with 4+ variants more than triple the expected impact.
- While high-variant testing is resource intensive and not always available, this shows that teams need to select the right tools for solving complex challenges.

More complex experiments yield greater returns

Optimizely's Web Experimentation allow for seven different types of changes to be mixed and matched: attributes, custom code, redirects, insert image, insert HTML, widgets, and CSS. While counting the number of different change types per test is not a perfect measure of complexity, it yields insights into a pattern we've long seen: complex tests outperform.

Variation expected impact by number of distinct code change types

Expected impact on checkout for true experiment variations, n = 18k variations



- Over half of all variations tested on checkout test one single change type.
- Variations with more changes offer greater expected impact on checkout.

Why does experiment complexity matter so much?

- **Low hanging fruit dries up**
  You can only change the color or text on a button so many times until you are at a local maximum. Gaining access to engineering resources to run more complex experiments is critical to maintain runway and continue your momentum.

- **Complex experiments move beyond cosmetic changes**
  Minute tweaks generally have minute effects on user behavior. To really change how a person interacts with a website or app, we need to tackle problems holistically and redesign experiences and journeys in a substantive way.

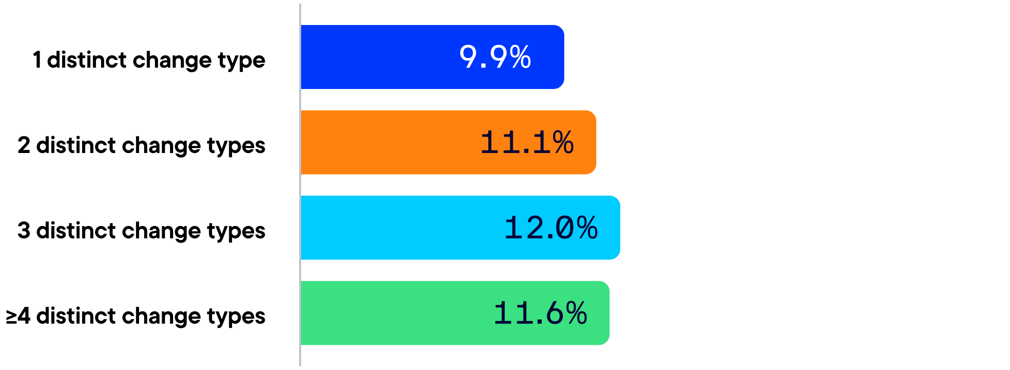
- **Complex experiments reflect ownership and responsibility**
  Programs that remain focused on minute optimization opportunities are often narrowly focused with limited freedom and resources. As programs gather more resources and gain more trust from the business, they receive the freedom to test more meaningful changes.

- **More time and effort is invested**
  Teams are unlikely to invest major effort in a complex experiment unless they feel more certainty about the value of the experiment.

Higher complexity variations outperform simple changes

Variation win rates by variation complexity

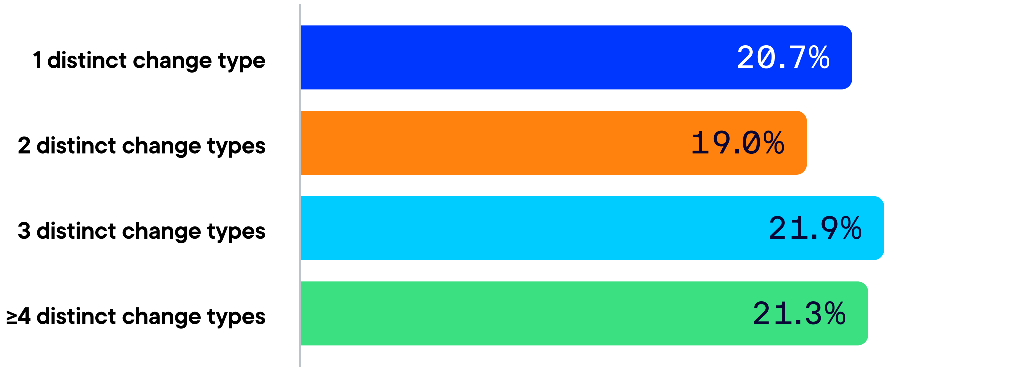
Win rates on primary metric for n = 128k true experiment variations on web



- Two-thirds of all variations run around the world test only one single change types.
- We see modest improvements in variation success rates as teams mix multiple types of changes within a single variant.

Average uplift for winning variations by variation complexity

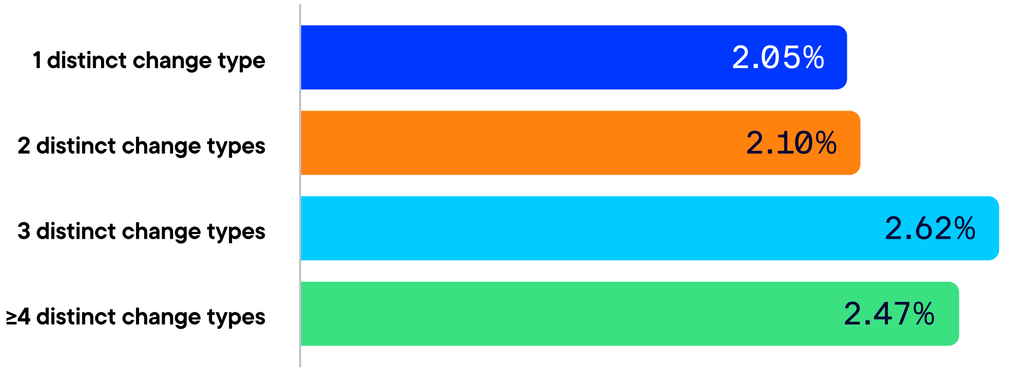
Avg. winning uplift on primary metric for n = 128k true experiment variations on web



Variation uplifts show a slight but positive increase when teams mix multiple types of changes.

Expected impact by variation complexity

Expected impact on the primary metric for n = 128k true experiment variations on web

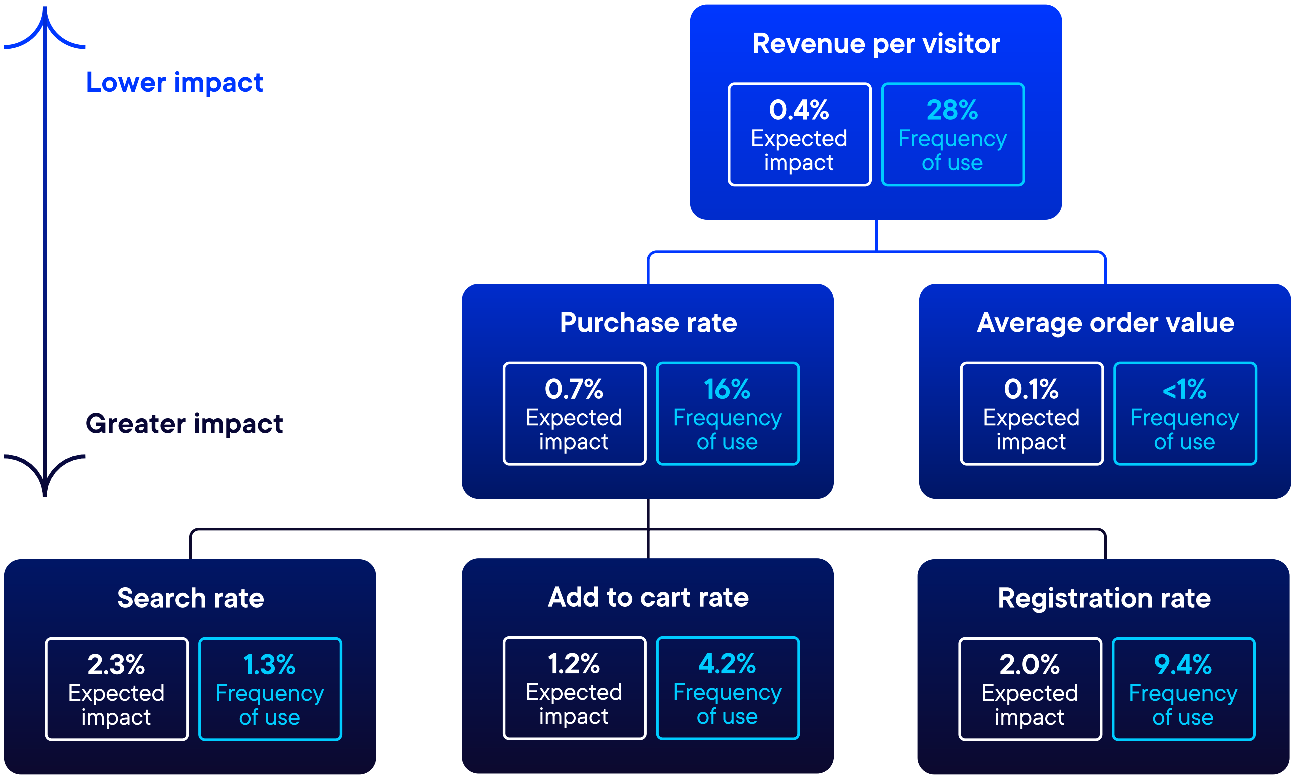


Expected impact of variations increases 25% as teams move from a single change type to 3.

Digital commerce has the highest returns from experiments that target goals which are early in the shopping funnel

> “Measuring every experiment on revenue is like measuring every player on points scored. Someone also needs to pass.”
>
> **Hazjier Pourkhalkhali**, Optimizely

Although revenue is the most valuable business metric, businesses stand to experience greater test impact by focusing experiments on improving micro-conversions, such as getting more users to search, add to cart, and register accounts.



Lower impact

Greater impact

Revenue per visitor

0.4% Expected impact

28% Frequency of use

Purchase rate

0.7% Expected impact

16% Frequency of use

Average order value

0.1% Expected impact

<1% Frequency of use

Search rate

2.3% Expected impact

1.3% Frequency of use

Add to cart rate

1.2% Expected impact

4.2% Frequency of use

Registration rate

2.0% Expected impact

9.4% Frequency of use

Search rate is the most undervalued experiment goal. Even though it is used 1% of the time, it has the highest expected impact at 2.3%. It is important to note that users who search typically convert at 2X-3X the conversion rate of all other users.

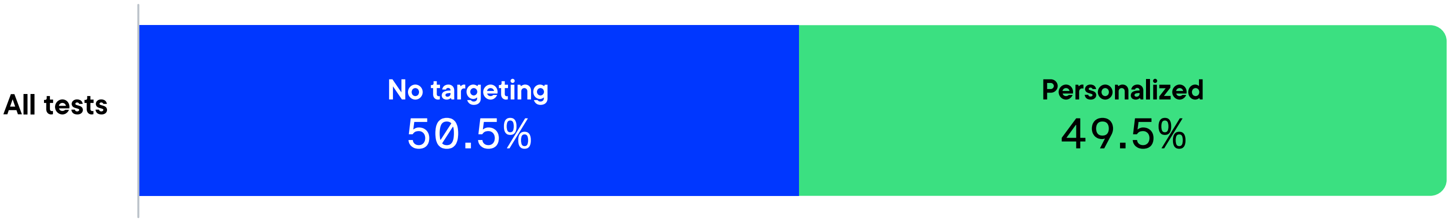
About half of experiments use audience conditions

In Optimizely, it is possible to target experiments to visitor groups using audience conditions; this can be considered as another dimension by which experiments can be made more complex. Common ways that audiences are defined include:

- Ad campaigns where a user originated
- Browser or version being used
- Cookie conditions such as whether a user is logged in
- Geolocation
- New vs returning users
- Traffic source
- Segments such as email subscribers, VIP members, etc.
- Any other attribute that is captured about the user

Half of all experiments run around the world use audience targeting

True experiments, n = 120k tests

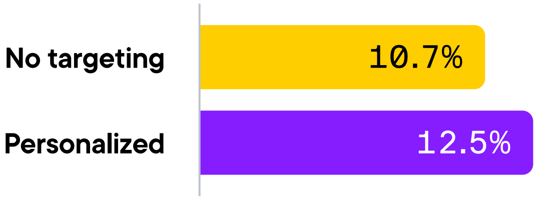


Just under half of experiments in Optimizely include audience conditions.

Personalized experiments can generate 41% higher impact on specific audiences than general experiences

Win rates by whether experiments use audience targeting

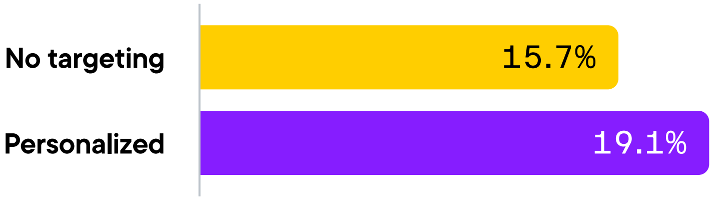
Win rates on primary metric for true experiments, n = 127k experiments



Experiments that include targeting are **16% more likely to win** when compared to untargeted experiment.

Average uplift by whether experiments use audience targeting

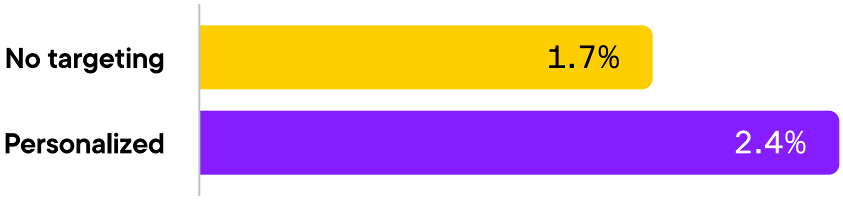
Average uplift on primary metric for true experiments, n = 127k experiments



Personalized experiences generate **22% higher uplifts** on average.

Expected impact by whether experiments use audience targeting

Expected impact on primary metric for true experiments, n = 127k experiments



- On a per-test basis, personalization results in **41% higher expected impact**.
- However, this impact will be mitigated by the reach of the audience.

04

Drivers of great cultures of experimentation

Great companies are built differently

Experimentation does not happen in a vacuum, and the resources and culture that support it are key to success. Data and analytics are key to formulate great hypotheses, and the right people are needed to create experiment variations.

Learnings from the most successful companies experimenting on Optimizely

Better tech stack and integrations fuel better experiments

Analytics capabilities and customer data platforms can be used to build data-driven hypotheses, and drive more successful experiments.

Various governance models can scale to high experimentation velocity

There is no one-size-fits-all approach when it comes to governance, organizations can be successful with any setup.

Having sufficient developer resources is crucial for success

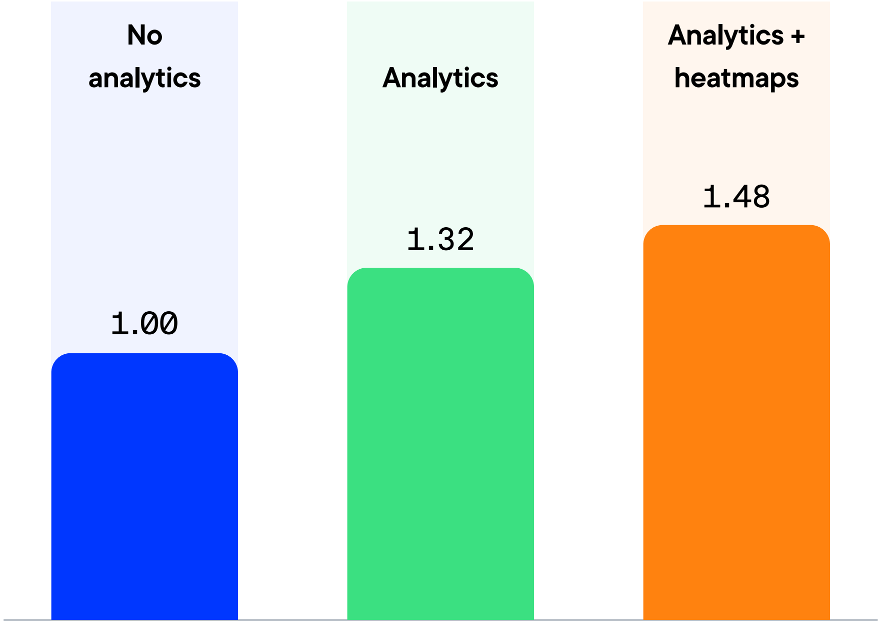
Stretching developers too thinly adversely affects experiment outcomes,
around one experiment per developer per sprint is the sweet spot.

Senior leadership can play an important role

Senior leaders can support experimentation by encouraging an innovative culture and ensuring sufficient resources.

Know your customers: companies with access to better analytics outperform those without

Experiment performance with or without analytics in the tech stack
Experiment win rate indexed against "No analytics", Optimizely + Built with data, 2018 n = >1,000 companies



- Teams with analytics outperform teams without by 32% per test.
- Teams that added heatmapping were an additional 16% more successful.
- Given that not all companies with analytics use it effectively, this suggests that analytics usage is a major improvement opportunity for more companies.

Great experimentation is based on effectively diagnosing and prioritizing user problems. Nonetheless, we find that many companies underinvest in analytics or, after purchasing a tool, insufficiently leverage data as a competitive advantage. Given the sizeable increases in performance seen above, companies should use all the resources they have access to and invest wisely.

Note: *This analysis is from 2018 and could not be replicated in 2023 because the number of companies without analytics has shrunk to near zero. Nonetheless, the insight remains as important as ever. Having data is not enough, you need to use it.*

> "The integration with the analytics tool allows KLM to automatically import experiment data for further analysis within a wider business context. Heatmaps can be automatically tagged with the information about the A/B test variation that a particular user has seen. This way the analysts can differentiate between experiences during their analysis."

KLM customer story

Customers with CDPs integrated with experimentation see higher win rates, uplifts and expected impact

↓

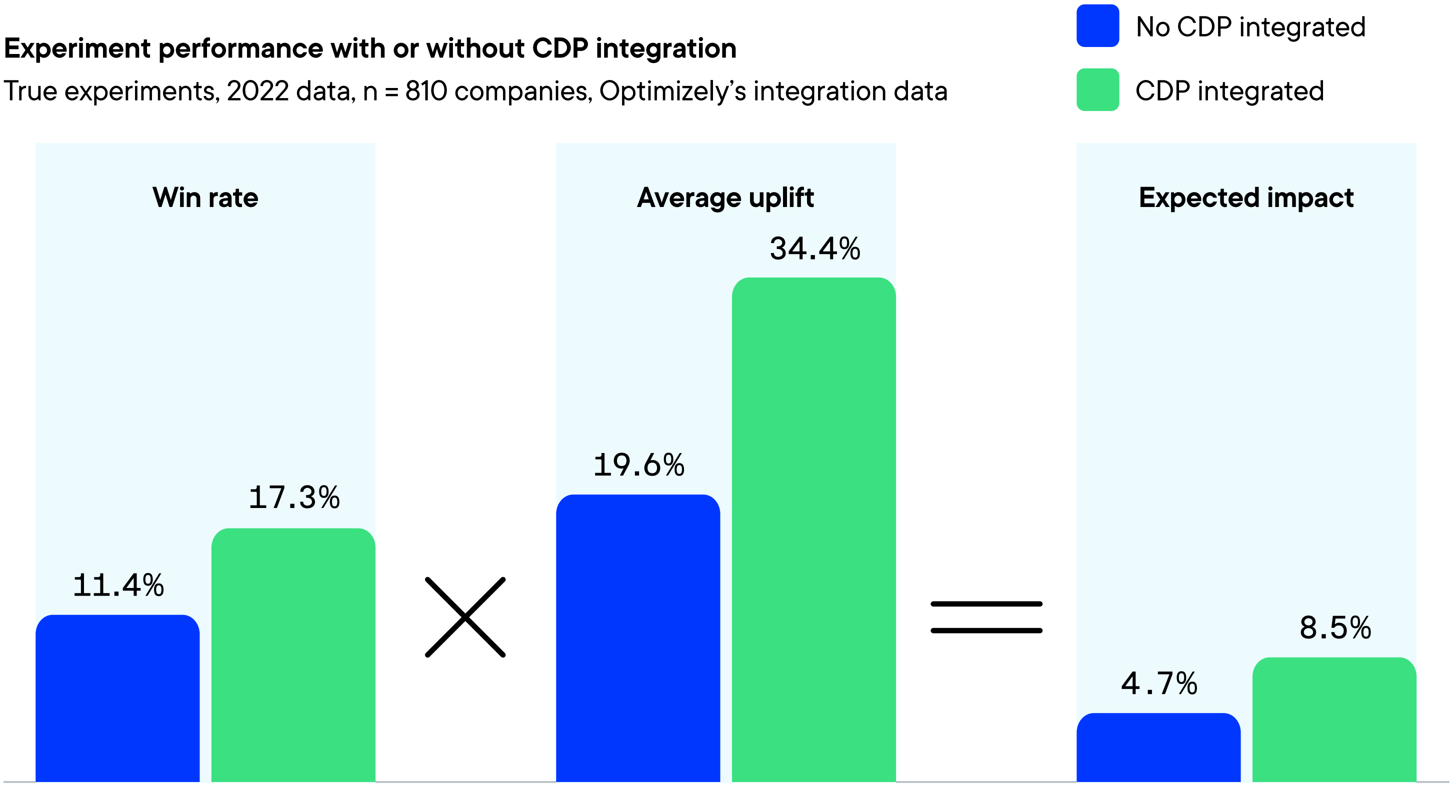
+80%

Expected impact

Companies who have a CDP and integrate it with Optimizely generate 80% greater experiment impact.

CDPs enable experimentation platforms to access a single source of customer data from your entire ecosystem.

Experiment performance with or without CDP integration
True experiments, 2022 data, n = 810 companies, Optimizely's integration data



Only around 7% of Optimizely customers with integrations have a CDP integrated with Optimizely's experimentation product, but this is associated with better experiment performance. There are likely confounding factors here as more digitally mature customers are more likely to have a CDP, but this data helps to highlight the need for a CDP as part of a digital maturity journey.

Advanced traffic allocation models such as Multi-Armed Bandits are underutilized

🟨 Original

🟧 Variation 1

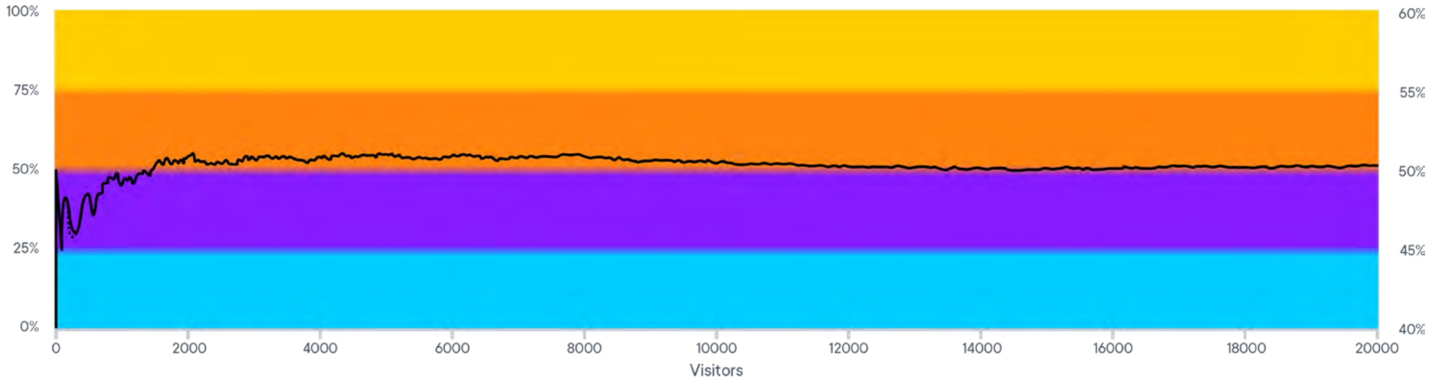
🟪 Variation 2

🟦 Variation 3

— Total conversion rate

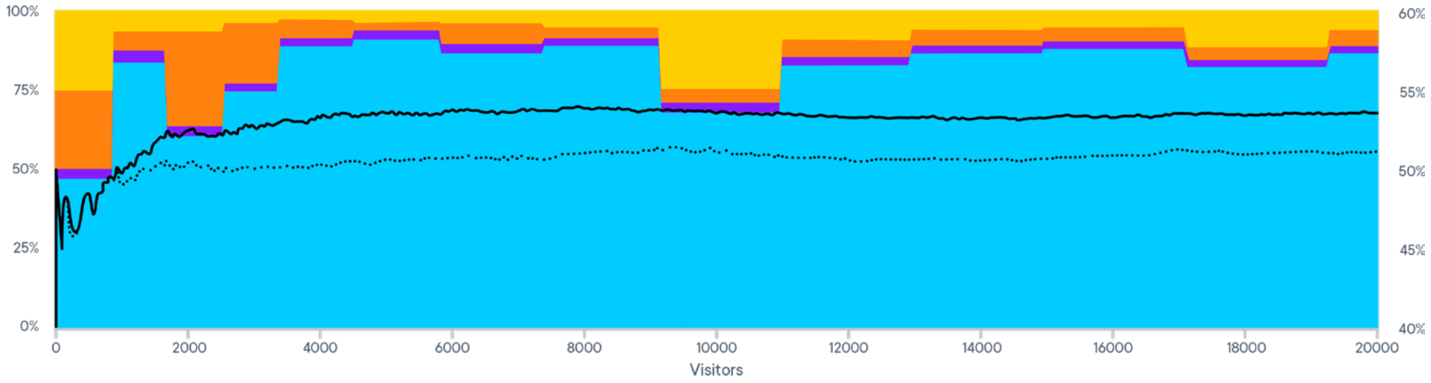
······· Counterfactual conversion rate

How normal traffic allocation looks



Standard A/B tests allocate traffic manually and keep a fixed split. Above, you can see four color-coded variations maintain the same 25% split throughout the entire test.

How Stats Accelerator / Multi-Armed Bandit allocate traffic



Optimizely's machine-learning algorithms continually reallocate traffic to improve one of two outcomes: faster significance (Stats Accelerator) or more conversions (Multi-Armed Bandit).

How stats accelerator reduces visitor needs for multi-variant tests

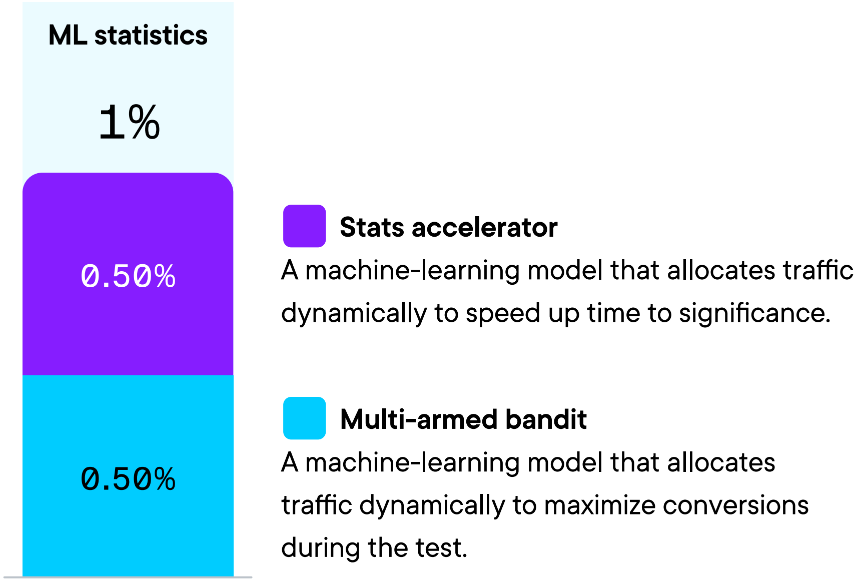
Optimizely research and simulations, 2019, assumes 100K visitors per variant

Number of variations	Manual allocation	Stats accelerator	Time to significance	Additional Tests / Year
2 variants	200K visitors	200K visitors	No benefit at 2 variants	No benefit at 2 variants
3 variants	300K visitors	230K visitors	23% faster	30% more
4 variants	400K visitors	260K visitors	35% faster	54% more
5 variants	500K visitors	290K visitors	42% faster	72% more

Tests with stats accelerator perform far better than manual traffic allocation

99% of all tests allocate traffic manually, only 1% via machine learning

Share of true experiments by traffic allocation model, n = 127k tests

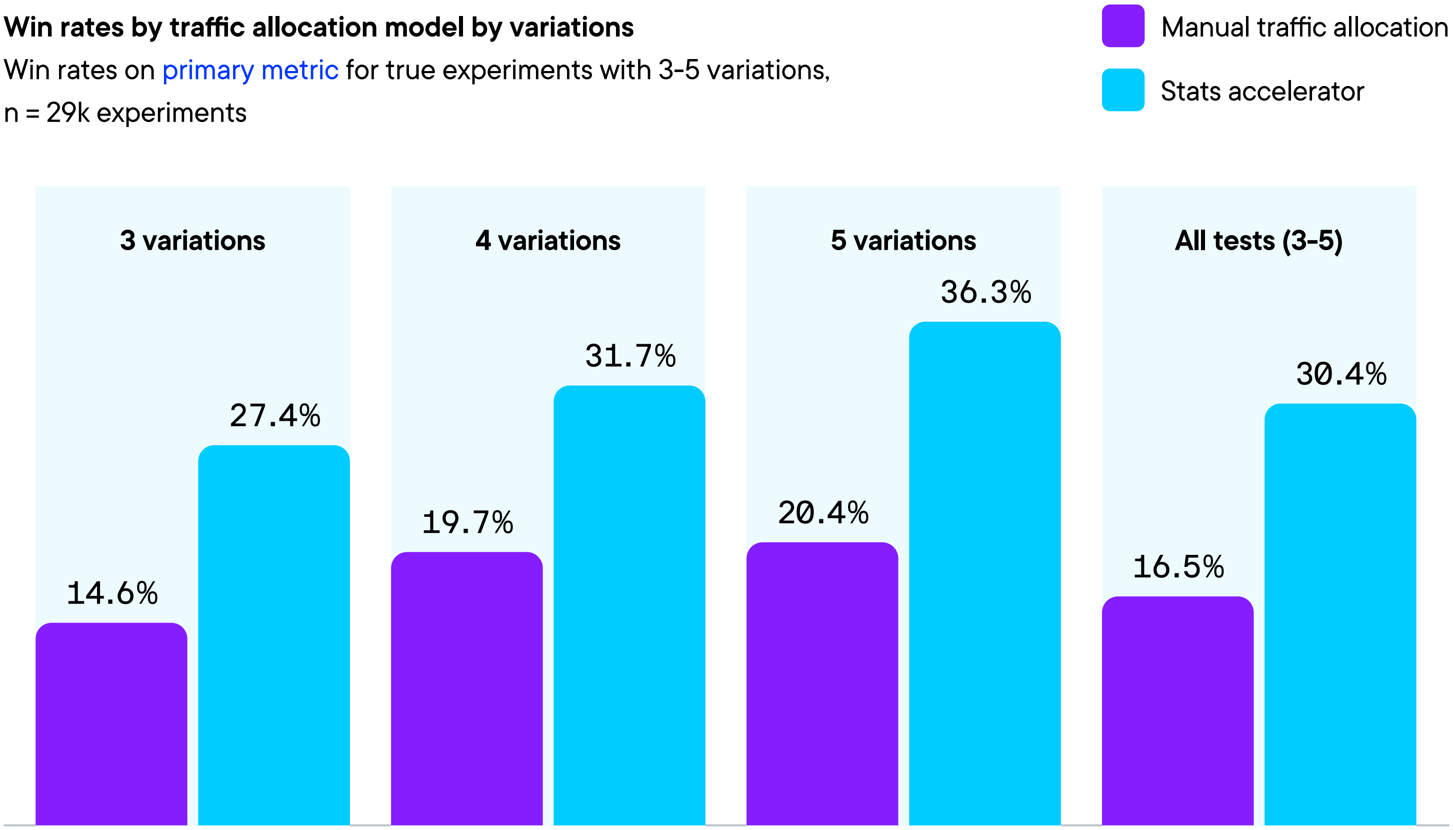


🟪 **Stats accelerator**
A machine-learning model that allocates traffic dynamically to speed up time to significance.

🟦 **Multi-armed bandit**
A machine-learning model that allocates traffic dynamically to maximize conversions during the test.

Win rates by traffic allocation model by variations

Win rates on primary metric for true experiments with 3-5 variations, n = 29k experiments



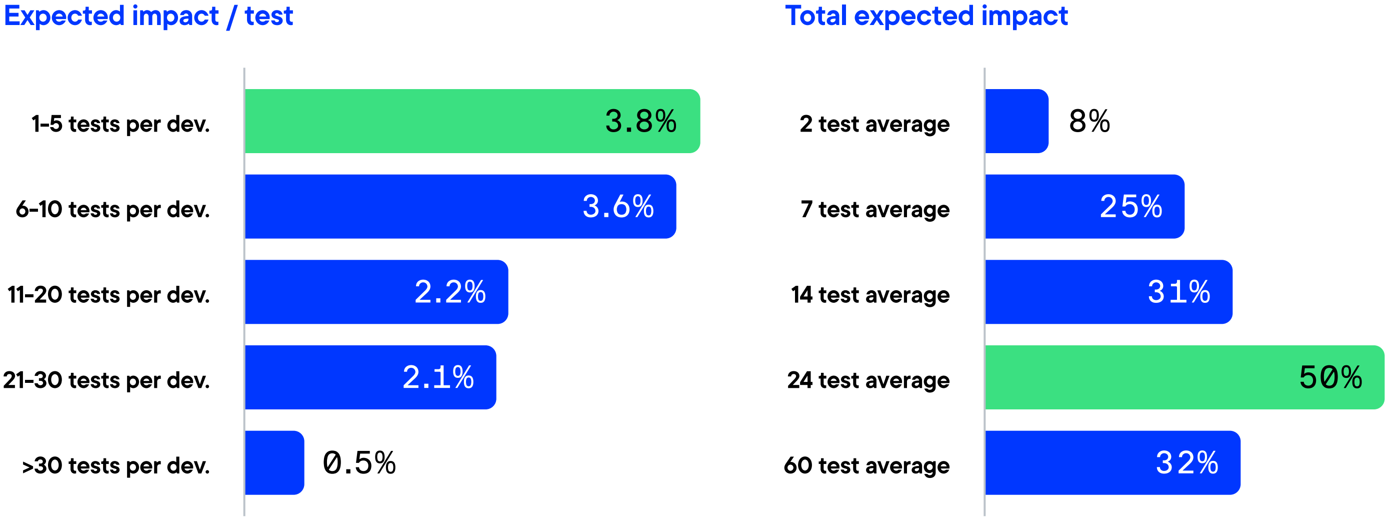
Experiments with stats accelerator win twice as due to three factors:

- **Time to significance**
  Stats accelerator requires only 30% of the traffic for each variation after 2; therefore, companies will win more often in less time.
- **Program maturity**
  Companies that use stats accelerator are more mature and have more sophisticated experimentation programs.
- **Past winners retests**
  Some companies retest prior best performing variations against each other with stats accelerator to select winners faster.

As a result, we do not believe the technology alone doubles success. The maturity of users and the context of when they use stats accelerator has a sizeable effect.

The most productive engineers run one experiment per two-week sprint cycle

Expected impact based on annual experiments per developer
True experiments run in 2022, n = 392 companies with 12+ experiments in 2022

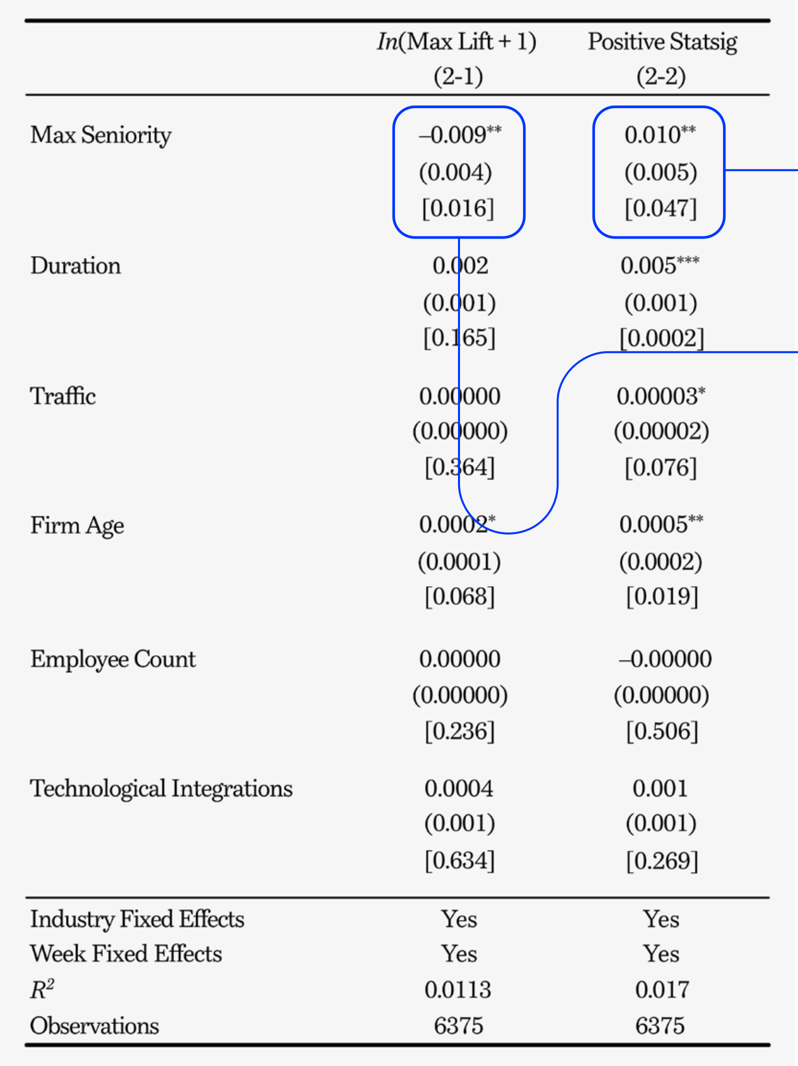


Engineering resources matter greatly

- **Teams with more engineers fare better**
  Companies often talk about the need for more testing velocity while they underinvest in developer resources. Without scaling engineering, experiment velocity risks becoming a vanity metric that worsens program outcomes.
- **The highest experiment quality occurs at 1-10 annual tests per engineer**
  Our data shows a 40% drop in expected impact per test once developers move to 11-30 tests per year, and an 87% drop in expected impact per test once developers move to >30 tests per year.
- **The highest productivity comes around 1 test per engineer per sprint**
  Despite a lower expected impact per test, the increase in velocity shows that engineers have their highest total impact when they are running around 24 tests per year, which is equal to one test per two-week sprint cycle per engineer.

Senior leaders associate with more winning ideas, yet junior teams associate with greater breakthroughs

Scientific study with Harvard Business School: Higher levels of seniority on testing teams associate with more winning experiments, yet smaller uplifts



	In(Max Lift + 1)	Positive Statsig
	(2-1)	(2-2)
Max Seniority	–0.009** (0.004) [0.016]	0.010** (0.005) [0.047]
Duration	0.002 (0.001) [0.165]	0.005*** (0.001) [0.0002]
Traffic	0.00000 (0.00000) [0.364]	0.00003* (0.00002) [0.076]
Firm Age	0.0002* (0.0001) [0.068]	0.0005** (0.0002) [0.019]
Employee Count	0.00000 (0.00000) [0.236]	–0.00000 (0.00000) [0.506]
Technological Integrations	0.0004 (0.001) [0.634]	0.001 (0.001) [0.269]
Industry Fixed Effects	Yes	Yes
Week Fixed Effects	Yes	Yes
R^2	0.0113	0.017
Observations	6375	6375

As the highest level of seniority found on a testing team rises...

1 Experiments appear to win more often.

2 Yet experiment uplifts are smaller than those of more junior teams.

This suggests that senior leaders have experience they can rely on to improve the status quo.

However, their known experience may close them off to more modern methods that can result in larger breakthroughs. Junior teams appear to take more risk, with fewer wins but greater uplifts.

Source: Ghosh, Sourobh, Stefan Thomke, and Hazjier Pourkhalkhali. “The Effects of Hierarchy on Learning and Performance in Business Experimentation.” Harvard Business School Working Paper, No. 20-081, February 2020.

> “Great leaders balance exploitation and exploration. They push teams to leverage and optimize business models that have worked in the past. But they also empower people to explore and discover new ways to create and capture value. Business experimentation is the engine that drives both endeavors.”

Prof. Stefan Thomke, William Clay Harding Professor of Business Administration at Harvard Business School and author of “Experimentation Works”

How senior leaders can contribute more effectively to experimentation programs

Common risks of seniority

Senior leaders can often overestimate their ability to influence the future, which closes them off to outside advice or critical feedback on projects.

Senior leaders are likely to use best practices that are moving out of date, causing them to focus on smaller improvement opportunities.

Senior leaders are less likely to revise their opinions when presented with data that conflicts with their beliefs than more junior team members.

Strong advantages of seniority

Senior leaders can effectively accelerate the adoption of new strategies and techniques through investments, strategies, guidance, and role modeling.

Senior leaders can increase their employees' psychological safety and freedom to take risks, which is known to improve performance.

Senior leaders can effectively balance exploitation and exploration, allowing teams to invest the right effort and take appropriate risks when called for.

Executive support is critical for any successful experimentation program. The most effective leaders focus on creating a system in which their employees can deliver their best work. This requires resourcing programs effectively, setting clear goals, ensuring career paths for key talent, opening doors, and building a culture of autonomy and open feedback.

“You need to ask yourself two big questions: How willing are you to be confronted every day by how wrong you are? And how much autonomy are you willing to give to the people who work for you? And if the answer is that you don’t like to be proven wrong and don’t want employees decide the future of your products, it’s not going to work.”

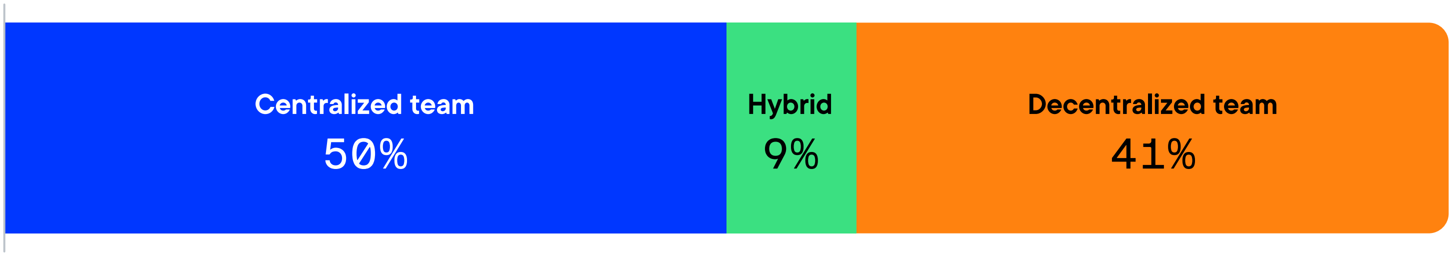
David Vismans, Chief Product Officer at Booking.com
Quoted in “Experimentation Works”

There is no one-size-fits-all governance model as companies report success with varying approaches

Governance models vary widely between companies and different times. Optimizely's interviews with customers over many years show governance models changing, regardless of low or high velocity.

A recent monthly customer survey corroborated these findings. While only a small sample, it directionally indicates what CRO experts have seen for years: there is no one-size fits all governance model. Companies must select the right model based on their team and business needs.

Who is responsible for running experiments at your company?
Optimizely customer survey, May 2023, n = 32 companies completing this question.



Most companies begin with a small group of experts of centralized CRO team.

Over time, these experts branch out and create more testing teams throughout the business.

At larger scales, companies are evenly split between centralized, decentralized, and hybrid models. Company approaches are based on the talent, business needs, and culture of the organization.

Best practices when scaling to multiple experimentation teams

- **Ensure effective collaboration**
  People need to share knowledge, develop scalable processes, document tools and templates.

- **Develop growth opportunities**
  There need to be career paths with clear expectations, opportunities for feedback, and access to coaching or mentorship.

- **Improve infrastructure**
  Teams need access to developers, high quality tooling, better analytics, good data governance, and validated experiment metrics.

- **Develop a culture of experimentation**
  The best programs decide through experiments and give employees freedom and safety.

Factors to consider when determining the right governance model for your business

1 **Control**

Who has the permission to publish an experiment? Who reviews the results and ultimately determines if a winner is implemented?

Ensure that other teams have learned the fundamentals of what makes a good experiment before opening the floodgates.

2 **Capabilities**

Do you have access to the right capabilities (and enough resources) to run complex experiments?

3 **Connection to the wider business**

Do you have a close relationship with the changing priorities of the core business? This is also essential for prioritization of your tests and growing your team.

Avoid remaining a siloed experimentation team. Ensure that your team is well connected to the business priorities and can pivot to provide that focus.

Common capabilities needed on experimentation teams

- Executive sponsor
- Program manager
- Technical lead
- Developer lead
- Content lead
- Analytics lead
- Design / UX

05 Concluding remarks

Concluding remarks

This report has confirmed insights that Optimizely uncovered in prior analyses. We're here to help other experimentation practitioners run better-quality experiments and increase the value they bring to their businesses.

Here are the biggest takeaways from the evolution of experimentation study:

1. **Resources to run complex experiments**
   That data shows more complex experiments outperform, but you need sufficient resources to start. They are two sides of the same coin. Both are important for success.

2. **Ideation and design**
   Grounding a hypothesis in data and measuring the right metrics impact how a practitioner performs ideation and design.

3. **Quantity vs. Quality**
   Data-driven experiments that focus on uplifts and impact drive more business value compared to simple, iterative experiments. Track and incentivize this behavior, instead of only maximizing throughput or your win rate in isolation.

Many Optimizely customers have already put these insights into practice and have seen outsized returns.

In the future, we will continue to further enrich these analyses. As AI capabilities improve, they can interpret the code attached to the experimentation data set and better categorize experiments. Plus, uncover meta-learnings around webpage design and UX.

For example:

- How do changes to calls-to-action on the homepage tend to perform?
- What are the characteristics of winning headlines?

As new features get added to the platform over time, it will be enlightening to determine how they can enhance the quality of experimentation programs; such features include our recently added automatic SRM detection and experimentation collaboration hub.

About the Authors

Elizabeth Gabster

Senior Director,
Strategy & Value

Elizabeth is a senior executive focusing on value, growth strategy, and experimentation. She has spent four years at Optimizely, leading Strategy & Value Consulting for Europe and previously spent four years consulting companies on their growth strategy at Google. Elizabeth is based in Amsterdam and received her MBA from INSEAD.

Eric Lang

Senior Consultant,
Strategy & Value

Eric focuses on helping Manufacturing & Distribution customers worldwide and Australia & New Zealand customers across all industries measure and maximize the returns of their commerce, experimentation, and digital experience programs. Prior to Optimizely, Eric managed processes, operations, and analytics in FinTech. Eric is based in Las Vegas and holds an MBA from the University of St. Gallen, Switzerland.

Kory Manley

Senior Director,
Strategy & Value

Kory has served as trusted advisor to the c-suite in industries ranging from private equity, retail, manufacturing, high tech and hospitality. Kory is based in Philadelphia and attended Ohio Northern University receiving an BA in Biology & Spanish and an ALM in Management from Harvard University. Kory has lived & worked in Spain, Cuba and Costa Rica.

Hazjier Pourkhalkhali

Global Vice President,
Strategy & Value

Hazjier has spent a decade in the field of experimentation at Optimizely across consulting, product, strategy, and value. He has run over 300 experiments and co-authored scientific research on experimentation with Harvard Business School. He previously consulted for McKinsey & Company. Hazjier is based in Amsterdam and studied Political Economics at UC Berkeley.

Emma Shillam

Lead Consultant,
Strategy & Value

Emma has spent 8 years in the field of experimentation and analytics across Optimizely and Mastercard (Applied Predictive Technologies), with a focus on digital experimentation and real-world in-store optimization. Emma is based in London and studied Economics at Cambridge University

Mark Wakelin

Lead Consultant,
Strategy & Value

Mark has spent the last 5 years working with organizations to help them advance their digital maturity. As a Senior Consultant in Optimizely's Strategy & Value team, he works with customers across Europe to help them identify and scale the value of their digital experience and experimentation programs. Mark is based in London and has an Msc in Strategic Marketing from Cranfield University.

06 Appendix

1/3

Definition 1: True experiments

The focus of our benchmark is to share trends and insights on experimentation. Therefore, other use cases of our products are excluded, which can include Feature Releases / Hotfixes, Personalization Rollouts, QA Environment Tests, and more.

Definition of a True experiment:

- There is a baseline variation with ≥ 1,000 visitors
- There are one or more non-baseline variations, with ≥ 1,000 visitors each and which are not A/A variants, measured as following:
  - Web experimentation: is there code attached to this specific variation
  - Feature experimentation: is the experiment not marked/named as A/A anywhere
- Experiment environment is 'Production' (not in 'Development', 'Staging', 'Demo', 'QA')

Example 1: True experiment on web experimentation

Variation	Visitors	Code change	Variant eligible?
Original	1,100	No	Yes
B variation	900	Yes	No
C variation	1,200	Yes	Yes
D variation	800	Yes	No
E variation	1,100	Yes	Yes
F variation	1,200	No	No

- True experiment, baseline and two non-baseline variant have ≥ 1,000 visitors and code changes.
- Three variations: Original + 2 Treatments.
  - B and D variations not recognized due to traffic.
  - F variant excluded as it is an A/A variant without code.

Example 2: Excluded experiment on feature experimentation

Variation	Visitors	Code change	Variant eligible?
Original	980	No	Yes
B variation	1,800	Yes	No
C variation	1,600	Yes	Yes

Words "A/A" not included in naming or test description

- **Excluded experiment**, baseline does not meet minimum traffic threshold. Test disqualified.
- Note **code change** is not inspectable in feature experimentation; however, there are no A/A test markers in the experiment or variation names.

2/3

Definition 2: Win / Lose / Inconclusive

Scientifically, experiments never win, they merely **disprove the null hypothesis**. However, in our industry, practitioners refer to experiments as “Winning”, “Losing”, or “Inconclusive.” To be more understandable, that same language is used here, with the terms defined below:

Definition of experiment outcomes:

Winning experiment

- True experiment
- Analyzes “Primary metric” unless specified otherwise
- For any variant, this metric has ≥ 90% significance and moves in the “Winning direction” (generally uplift)

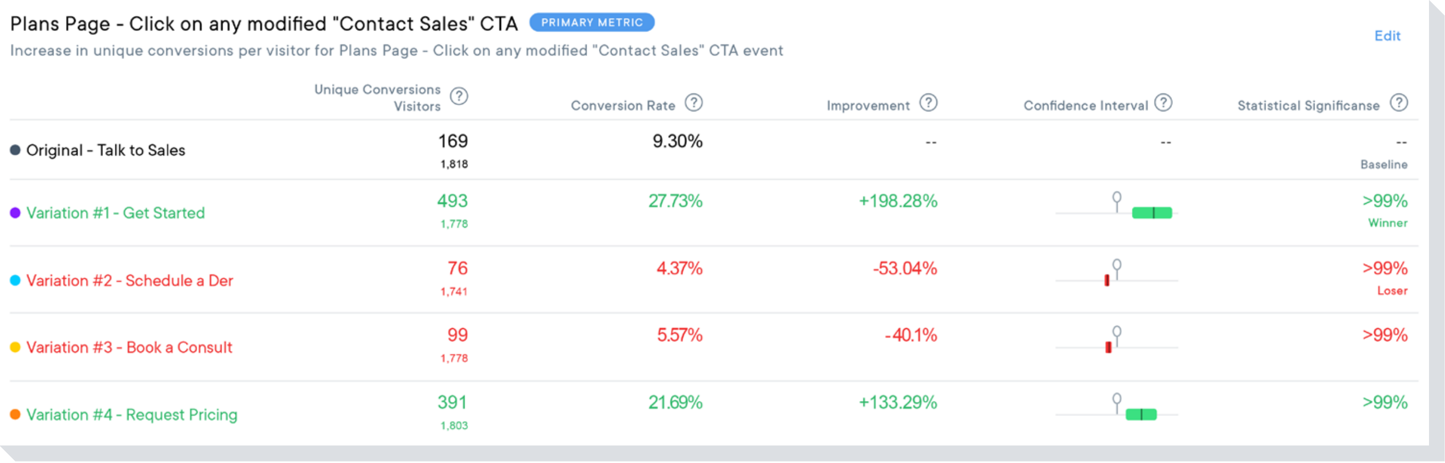
Losing experiment

- True experiment
- Analyzes “Primary metric” unless specified otherwise
- For any variant, this metric has ≥ 90% significance, and moves in the “Losing direction” (generally a reduction)
- No variant is a winner

Inconclusive experiment

- True experiment
- Analyzes “Primary metric” unless specified otherwise
- No variant achieves ≥ 90% significance

Real example: Experiment run by Optimizely on our plans page



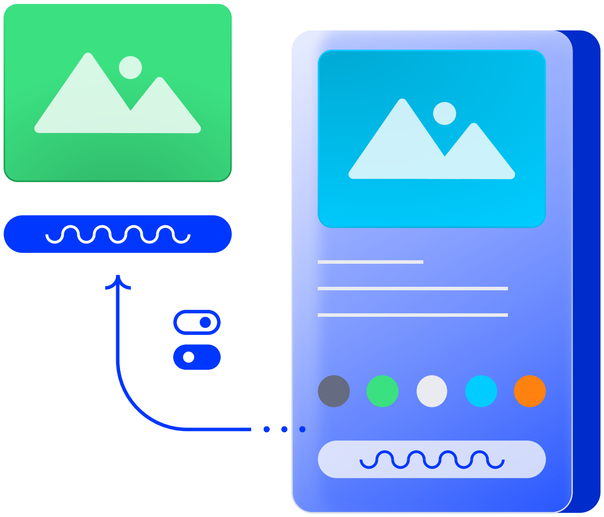
Plans Page - Click on any modified "Contact Sales" CTA PRIMARY METRIC

Increase in unique conversions per visitor for Plans Page - Click on any modified "Contact Sales" CTA event

Edit

	Unique Conversions Visitors ⓘ	Conversion Rate ⓘ	Improvement ⓘ	Confidence Interval ⓘ	Statistical Significanse ⓘ
● Original - Talk to Sales	169 1,818	9.30%	--	--	-- Baseline
● Variation #1 - Get Started	493 1,778	27.73%	+198.28%		>99% Winner
● Variation #2 - Schedule a Der	76 1,741	4.37%	-53.04%		>99% Loser
● Variation #3 - Book a Consult	99 1,778	5.57%	-40.1%		>99%
● Variation #4 - Request Pricing	391 1,803	21.69%	+133.29%		>99%

- **True experiment**, sincere there is a Baseline with ≥ 1,000 visitors, and all treatments have ≥ 1,000 visitors and make a change; in this case, modifying wording. No variations are dropped.
- **Winning experiment** as "Variation #1 - Get Started" has statistical significance and moves in the Winning Direction on the Primary Metric, "Contact Sales." Even though Variations #2 and #3 are losing, practitioners will always implement the best, not worst, variants; therefore, "Winning" always overrides "Losing."



3/3

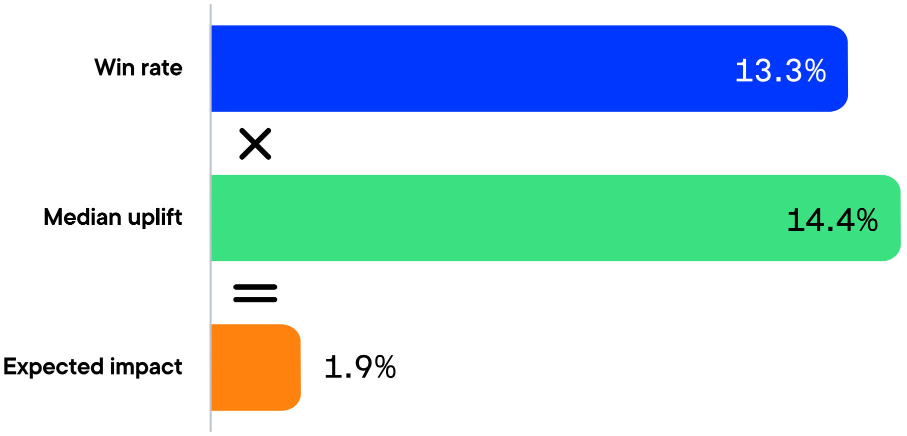
Definition 3: Expected impact

In our view, ”winning experiments” are critical, but without context, they risk becoming a vanity metric or distraction. There are many companies with high “win rates” but very low uplifts, since they pursue minute, low risk changes and bug fixes. There are also companies that have very low win rates for their program, because they take sizeable risks that result in major winners when the experiments succeed.

In order to take a more holistic view, we increasingly analyze companies by the expected impact of their experiments, defined as: how likely is an experiment to win times what is the average uplift for winning experiments. Below we share the Expected Impact calculation for the pageviews metric.

Experiment outcomes on pageviews

n = 33,210 True experiments, for the “pageviews” metric.



What percentage of experiments have at least one winning variation for this metric?

What is the average for all winning experiments of their best performing variation’s uplift?

For each test that is run, what return is expected given the probability and value of winning?

Advice to senior leaders

A number of programs try to drive up performance by pushing the team for more winning tests: “raise the velocity and win more often”.

But some teams react by running a high volume of bug fixes and obvious changes as experiments. Velocity improves, win rates go up, but average uplift drops. As a result, the program now produces less value than previously. To steer performance more effectively, there is a need to focus on more holistic metrics, such as expected impact.

At Optimizely, we're on a mission to help people unlock their digital potential. We do that by reinventing how marketing and product teams work to create and optimize digital experiences across all channels. With our leading digital experience platform (DXP), we help companies around the world orchestrate their entire content lifecycle, monetize every digital experience, and experiment across all customer touchpoints. Optimizely has 700+ partners and nearly 1500 employees across our 21 global offices. We are proud to help more than 10,000 businesses, including H&M, PayPal, Zoom, Toyota, and Vodafone, enrich their customer lifetime value, increase revenue, and grow their brands. At Optimizely, we live each day with a simple philosophy: large enough to serve, small enough to care. Learn more at **Optimizely.com**