CONTENTS

FOR DTC FOUNDERS AND OPERATORS · A FIELD GUIDE

THE HONEST TEST

Conversion work for stores without Amazon's traffic, and without room to fool themselves.

ANDREW LAUCHNER Author of The Second Order and The Whole Machine

September 2026 · 15 chapters · About an hour

A note before you start

Most DTC stores run A/B tests they can't read. The tool shows a winner, the team ships it, and six months later the conversion rate is where it started. Nobody lied. The math was never there.

8%	~1,600
of ideas tested in Airbnb's search experiments improved the metric they targeted: 20 of 250 (Ron Kohavi)	orders in each version of a test, at the usual settings, to detect a 10% lift in conversion rate reliably

Those two numbers are the guide in miniature. The first says that at a company with some of the best product people and data in the world, most ideas didn't work. Most of yours won't either, and that's normal. The second says what it costs to find out: about 1,600 orders in each version to see a 10% lift, at any typical store conversion rate. A store taking 400 orders a month through its product pages would need eight months to run that test once.

So this is a guide for the rest of us. It shows the math the testing tools don't lead with, and then what to do when the math says you can't test: which research finds the leaks, which fixes don't need a test at all, and how to change a product page with confidence when a clean experiment isn't available.

> **A result you can't read is not a result. It's a decision you made anyway, with a chart attached.**

It builds on The Whole Machine, whose chapters on the product page and on tests are a shorter version of parts one and three here.

How to read it

Start with The Testing Audit, or with the traffic table just below, which tells you in one look what your traffic can support. Or follow a path:

- **Founder:** Your Traffic on One Page, Most Ideas Don't Work, The Ladder of Evidence, Fix It, Don't Test It.
- **Running tests already:** How Much Traffic a Test Needs, Why Winners Lie, Before You Believe a Result.

- **Designer or site lead:** Research That Finds the Leak, The Product Page, The Mobile Gap.
- **Analyst:** part one, then For Your Analyst.

Four tools run in the page. Nothing you type leaves your browser.

What's proven and what isn't

- **Published:** peer-reviewed research, books and papers by the researchers themselves, official documentation and law, with the date it was true.
- **Reported:** what a named company, research firm or outlet published. Testing-tool companies and analytics vendors publish much of the data in this field, and they sell the thing they measure, so the text names the vendor each time.
- **Derived:** my arithmetic, using a standard formula, with the inputs shown.
- **From my work:** methods I use on client stores. Clients aren't named and their figures aren't here.

Examples that open with Say or Picture use made-up round numbers. Every source is listed in Appendix C.

Andrew Lauchner

Scottsdale, Arizona

FRONT

TEN POSITIONS

What this guide argues, and what would prove each claim wrong.

A position says what would prove it wrong. Test each on your own store.

01 **Most ideas don’t move the number they target, even at the best companies.**

Wrong if more than half of your fully powered tests produced a win that held up when you tested it again.

Most Ideas Don’t Work

02 **Traffic decides what you can test, before any idea does.**

Wrong if your store can detect a 5% lift on its product pages in a month.

How Much Traffic a Test Needs

03 **A winner from an underpowered test is more likely to be noise than a winner from a well-powered one, and its lift is exaggerated either way.**

Wrong if your winners’ lifts hold at full size when you run them again.

Why Winners Lie

04 **Stopping a test when it looks good multiplies false winners.**

Wrong if, across many A/A tests checked daily and stopped at the first significant result, no more than one in twenty finds a winner.

Why Winners Lie

05 **Check the split before the result.**

Wrong if none of your past tests had a traffic split off by more than chance allows.

Before You Believe a Result

06 **When you can’t test, research tells you what to change and a careful before-and-after tells you whether it helped.**

Wrong if five customer sessions on your product page find nothing you didn't already know.
The Ladder of Evidence

07 **Bugs, slowness and missing information get fixed, not tested.**

Wrong if fixing known checkout bugs has, on balance, lowered your checkout completion rate.
Fix It, Don't Test It

08 **Most product pages lose sales by hiding answers: the total cost, the delivery date, the returns, the reviews.**

Wrong if your post-purchase survey's "what almost stopped you" answers are about something else.
The Product Page

09 **Judge conversion changes on contribution, not conversion rate.**

Wrong if your free-shipping threshold or discount test raised conversion and margin every time.
Shipping, Offers and Price Tests

10 **A test that wasn't written down before it ran teaches nothing, even when it wins.**

Wrong if your team can say what it learned from its last ten tests without looking them up.
The Brief and the Log

FRONT

YOUR TRAFFIC ON ONE PAGE

Find your row. It tells you the smallest improvement a four-week test on that page could reliably see, and so which kind of evidence your store can afford.

Count the visitors who reach the page you want to change, not the site's total. Then read across.

VISITORS TO THE PAGE IN 4 WEEKS	SMALLEST LIFT A 4-WEEK TEST CAN RELIABLY DETECT, AT 2% CONVERSION	AT 3% CONVERSION	WEEKS TO SEE A 10% LIFT, AT 2%
10,000	40%	32%	63
25,000	25%	20%	26
50,000	18%	14%	13
100,000	13%	10%	7
250,000	8%	6%	3
1,000,000	4%	3%	1

DERIVED Two versions with traffic split evenly, a two-sided 5% significance level and 80% power, using the standard rule of thumb for sample size in chapter 3. Lifts are relative: a 20% lift takes 2.0% to 2.4%. Weeks are rounded up to whole weeks.

Now hold that against what real changes do. At Bing, one of the most tested products in the world, a change to how ad headlines were displayed raised revenue by 12%, worth over $100 million a year in the US; the book that tells the story says simple changes that big happen there "only once every few years" **PUBLISHED** . If your row says the smallest lift you can see is 18% or more, a clean A/B test can't reliably see most of the changes you're likely to make. It will still show you winners. They'll mostly be noise.

> **Your traffic decides what you can learn from a test before your ideas do.**

What each row can do

- **Under about 50,000 visitors to the page in four weeks:** don't A/B test small changes. Use research to decide, fix what's broken, and read changes with the before-and-after method in chapter 6. Test only big swings: a different offer, a different page, a different price.
- **About 50,000 to 250,000:** test bold changes on your highest-traffic pages, one at a time, and run them to the planned size. Everything else goes through research and fixing.
- **Above about 250,000:** you can run a real testing program. Part one is about running it without fooling yourself.

Do this

- **Pull four weeks of visitors and conversion rate** for your three most important pages: usually the home page, the top product page and the cart. Find each one's row. Most teams discover that only one page, if any, can carry a test.

START HERE · CHAPTER 1

THE TESTING AUDIT

Twelve checks on whether your store can learn from changes, and whether it has been. About forty minutes with your analytics, your testing tool and a phone.

The audit isn't about how many tests you run. It's about whether the results you act on are true, and whether the changes you make without a test are chosen by evidence or by whoever spoke last in the meeting.

Open your analytics, your testing tool if you have one, your store's checkout settings, and your last ten tests or site changes. Score each check 0 to 2: 0 if it failed or nobody can answer it, 1 if partly true, 2 if clean. A tool's "95% chance to win" doesn't count as an answer to any of these.

> **A dashboard's "chance to win" is a claim, not a check.**

The twelve checks

---

01 **You know each key page's traffic and conversion rate · 4 min**

Look at: Four weeks of visitors and conversion rate for your home page, top product page and cart, by device.
Good: Someone can produce the six numbers today, and knows which row of the traffic table each page sits in.
Cost if wrong: You start tests that can't finish, and believe the ones that stop early.
Read next: How Much Traffic a Test Needs

---

02 **Every test has a planned size before it starts · 3 min**

Look at: Your last five tests.
Good: Each had a sample size and end date written down before launch, based on the smallest lift worth acting on.
Cost if wrong: Tests end when someone gets bored or excited, which means they stop when the noise looks best.
Read next: How Much Traffic a Test Needs

---

03 **One primary metric, chosen in advance · 3 min**

Look at: The same five tests.
Good: Each named one metric that would decide it, before it ran. Usually orders or contribution per visitor.

Cost if wrong: With enough metrics, every test finds a winner somewhere.
Read next: Why Winners Lie

---

04 **Nobody stops a test early because it looks good · 3 min**

Look at: Whether any recent test ended before its planned size.
Good: None did, or the tool uses a method built for checking results as they arrive, and the team knows which.
Cost if wrong: Peeking can turn a 5% false-winner rate into 25% or more.
Read next: Why Winners Lie

---

05 **Winners are confirmed · 3 min**

Look at: What happened after your last three winners shipped.
Good: Each was re-run, held back from a slice of traffic, or checked against the forecast after launch.
Cost if wrong: You bank lifts that were never there.
Read next: Why Winners Lie

---

06 **The split is checked on every test · 3 min**

Look at: Visitors in each version of your last three tests.
Good: Someone checked the split was as planned before reading the result.
Cost if wrong: A broken split invalidates the result, and it happens more often than teams expect.
Read next: Before You Believe a Result

---

07 **An A/A test in the last year · 2 min**

Look at: Whether you've ever run two identical versions against each other on your testing tool.
Good: Yes, and the split was even and no winner held up on a rerun.
Cost if wrong: You're trusting a setup nobody has checked.
Read next: Before You Believe a Result

---

08 **Research in the last quarter · 4 min**

Look at: Recordings or notes from customer sessions, and your post-purchase survey.
Good: You watched at least five people use your site on a phone, and a live post-purchase question asks buyers what almost stopped them.
Cost if wrong: Your test ideas come from opinions instead of from customers.
Read next: Research That Finds the Leak

---

09 **Known problems are fixed, not tested · 4 min**

Look at: Your list of known bugs, slow pages and errors.
Good: There's a list, it's short, and nothing on it is waiting for a test.
Cost if wrong: You spend traffic proving a broken thing is broken.
Read next: Fix It, Don't Test It

10 **The product page answers the four questions · 5 min**

Look at: Your top product page, on a phone.
Good: Total cost, delivery date, returns and reviews visible near the add-to-cart button, without scrolling far or opening anything.
Cost if wrong: Shoppers find the answers at checkout, and about 70% of carts are abandoned before an order.
Read next: The Product Page

11 **Mobile has its own number and an owner · 3 min**

Look at: Your weekly report.
Good: Mobile conversion rate is reported separately from desktop, and one person owns closing the gap.
Cost if wrong: The device that brings most of your traffic gets the least attention.
Read next: The Mobile Gap

12 **A test log · 3 min**

Look at: Where past tests and site changes are recorded.
Good: One searchable log, with the hypothesis, the planned size, the result and what changed because of it, for every test and every major change.
Cost if wrong: You'll run the same losing test again next year.
Read next: The Brief and the Log

Score as you go; your band appears when all twelve are in.

RUN YOUR NUMBERS

Score the twelve checks

0: failed, or nobody can answer it. 1: partly true. 2: clean. Scores stay in this browser.

01 You know each key page's traffic and conversion rate 4 min	0	1	2

02 Every test has a planned size before it starts 3 min	0	1	2
03 One primary metric, chosen in advance 3 min	0	1	2
04 Nobody stops a test early because it looks good 3 min	0	1	2
05 Winners are confirmed 3 min	0	1	2
06 The split is checked on every test 3 min	0	1	2
07 An A/A test in the last year 2 min	0	1	2
08 Research in the last quarter 4 min	0	1	2
09 Known problems are fixed, not tested 4 min	0	1	2
10 The product page answers the four questions 5 min	0	1	2
11 Mobile has its own number and an owner 3 min	0	1	2
12 A test log 3 min	0	1	2

0 of 24 points	0 of 12 checks scored

Score all twelve to read your band. 12 to go.

RESET SCORES

Read your score

SCORE	WHAT IT MEANS	READ NEXT
20–24	You can trust what you learn. Your job now is learning faster: bolder tests on the pages that can carry them.	What One Good Test Looks Like, then The Brief and the Log
14–19	Some of what you believe is true and some isn't, and you can't yet tell which. Fix the zeros first.	The chapter linked from your lowest check, then Why Winners Lie
8–13	You're changing the site on opinion, with a testing tool to make it look like evidence.	Part one, starting at How Much Traffic a Test Needs
0–7	Stop testing for a month. Fix what's broken and talk to five customers.	Fix It, Don't Test It, then The First Thirty Days

If you've never run a test, checks 2 to 7 score 0 and that's fine: they're about running tests well. Ignore the band and look at checks 1 and 8 to 12, which apply to every store. A clean 12 on those six is a strong start.

PART ONE · THE MATH NOBODY SHOWS YOU · CHAPTER 2

MOST IDEAS DON'T WORK

At the best-run testing programs in the world, somewhere between one idea in three and one in twelve improves the number it was meant to improve. Plan for that.

Every test starts with someone who believes in the idea. Ron Kohavi, who ran experimentation at Microsoft and later at Airbnb, spent two decades counting how often that belief is right. The answer is the most useful number in conversion work.

The rates

WHERE	SHARE OF TESTED IDEAS THAT IMPROVED THEIR TARGET METRIC
Microsoft	About one in three
Bing	About 10% to 20%
Booking.com	About 10%
Google (2009)	About 10% of roughly 12,000 experiments led to a change being made
Airbnb search	8%: 20 of 250 ideas

PUBLISHED Kohavi, Crook and Longbotham, 2009; Kohavi and colleagues, 2012, 2014 and 2022, the last citing Stefan Thomke for Booking.com and Jim Manzi for Google. Full references in Appendix C.

These are products tuned by thousands of engineers, so the easy wins are long gone, and a small DTC store can reasonably expect a higher hit rate on its first ideas. But the pattern holds: most ideas won't work. At Microsoft, the two-thirds of ideas that didn't improve their metric either made no measurable difference or made things worse. Nobody could reliably tell which ideas would work before testing them.

Nobody can pick the winners

The clearest example comes from Bing. An engineer suggested showing more of an ad's text in its headline. It was a small change, rated as low priority, and it sat in the backlog for more than six months. When someone finally tested it, revenue rose 12%, worth more than $100 million a year in the US PUBLISHED . Kohavi and Stefan Thomke later called it the best revenue-generating idea in Bing's history. Nobody had thought it would matter.

> **Your confidence in an idea tells you almost nothing about whether it will work. That's why you test, and why the test has to be one you can believe.**

What this means for a small store

- **Expect most tests to lose.** If more than half your tests are "winning," something is wrong with how you're reading them. Chapter 4 covers what.
- **A loss is worth something.** Every change you didn't ship because a test caught it is a small disaster avoided. Two-thirds of Microsoft's ideas would have shipped on opinion alone.
- **Where ideas come from matters more than how many you test.** If you can only run a handful of real tests a year, the hit rate of your ideas decides whether you find anything. In my experience, ideas drawn from watching customers struggle win more often than ideas copied from a competitor's homepage. Chapter 7 is about finding them.

Do this

- **List your last ten tests or major site changes** and mark each as won, lost or no difference. If more than half won, keep that list open while you read the next three chapters.

PART ONE · THE MATH NOBODY SHOWS YOU · CHAPTER 3

HOW MUCH TRAFFIC A TEST NEEDS

One formula, sixteen times a number over a number squared, tells you before you start whether a test can find what you're looking for.

Testing tools will happily start any test. None of them leads with the question that decides whether it's worth starting: how many visitors does it take to see the lift you care about? The answer comes from a rule of thumb that statisticians have used for decades.

The rule

For a test with two versions and traffic split evenly, the number of visitors each version needs is about:

VISITORS PER VERSION

$$16 \times p \times (1 - p) \div d^2$$

p is the conversion rate now. d is the change you want to detect, in absolute terms: to see a 10% lift on a 2% conversion rate, d is 0.2 percentage points, or 0.002.

This is Lehr's rule, and Kohavi and his co-authors give it for exactly this use: it provides a two-sided 5% significance level and 80% power, which means that if the lift is real and that size, the test will detect it four times in five. Replace the 16 with 21 for 90% power PUBLISHED .

Work it through and a shortcut appears. To see a 10% lift you need about 1,600 orders in each version, and to see a 5% lift about 6,300, almost regardless of your conversion rate DERIVED . The statistician Martin Goodson, then at the testing company Qubit, gave the same rule of thumb in 2014: 1,600 conversions per group for a 10% lift, 6,000 for 5% REPORTED . So the fastest way to know if you can test a page is to count its orders. A product page that produces 400 orders a month needs about eight months to see a 10% lift.

> **Count the orders, not the visitors. About 1,600 per version to see a 10% lift.**

Kohavi's own summary is blunt: A/B tests are useful for effects of reasonable size when you have "at least, thousands of active users, preferably tens of thousands" PUBLISHED . Compare that with what some tools will accept. VWO, a popular testing tool, says its engine needs at least 25 conversions per version and 1,500 visitors in total before it will call a result REPORTED . That's enough to show a winner. It's nowhere near enough to know if the winner is real.

Revenue needs more than orders

Conversion rate is the cheapest metric to test, because every visitor either orders or doesn't. Revenue per visitor is far noisier, because a few large orders swing it. In Kohavi's worked example, detecting a 5% change in revenue per user took 3.3 times as many users as detecting a 5% change in conversion rate PUBLISHED . In studies of retail advertising, a customer's sales commonly varied by ten times their average PUBLISHED . So test on orders per visitor, and check that order value didn't fall. Changes that aim straight at order value, like bundles, price points and free-shipping thresholds, need revenue or contribution as the metric, and that needs a big store. Chapter 11 has the alternatives.

Test where the traffic is

The visitors that count are the ones who see the change. A new product page layout is seen by product page visitors, not by the whole site. A checkout change is seen only by people

who reach checkout, a small fraction of your traffic but with a high conversion rate, which helps. Run the numbers for the page, and prefer changes that apply across many pages at once, such as a template, where the traffic adds up.

RUN YOUR NUMBERS

Can this page carry a test?

Example numbers. Replace with yours. Visitors means people who reach the page you'd change, both versions together.

Visitors to the page per week: 6000

Their conversion rate now: 2.5 %

Smallest lift worth acting on: 10 %

Longest you'd run a test: 4 weeks

Yearly revenue through this page: $ 1200000

62,400 visitors needed in each version, about 1,560 orders each	21 weeks to run, at your traffic	22.8% smallest lift you can reliably see in 4 weeks	24% chance of detecting a real 10% lift if you stop at 4 weeks

$120,000

extra revenue a year, if the lift is real (before costs)

> **A 10% lift would take 21 weeks to see. In 4 weeks you can only reliably see lifts of 22.8% or more, so test only bold changes here, or use the ladder in chapter 6. Stopping at 4 weeks gives a real 10% lift a 24% chance of showing up.**

Two versions, traffic split evenly, a two-sided 5% significance level, 80% power. Round the result up to whole weeks, and run whole weeks: weekday and weekend shoppers behave differently.

Look at the fourth number in the example. A test stopped at four weeks, on a page that needed twenty-one, has about a one-in-four chance of detecting a real 10% lift. Put that power into the tool in the next chapter and see what it does to the winners.

Do this

- **Run the planner for your top three pages** at a 10% lift. Write the weeks needed next to each page on your test roadmap. Any test planned for a page that needs more than your longest acceptable run (four weeks is a good default; past eight, expiring cookies start to mix the groups) either becomes a bolder change or leaves the roadmap.

PART ONE · THE MATH NOBODY SHOWS YOU · CHAPTER 4

WHY WINNERS LIE

A statistically significant winner can still be false, and a real winner is usually smaller than it looked. Three habits make both worse.

“Significant at 95%” sounds like “right 95% of the time.” It isn’t. It means that if the change did nothing, a difference at least this big, in either direction, would show up by chance less than one time in twenty. How often a winner is real depends on something the p-value doesn’t know: how many of your ideas work in the first place.

False winners

Kohavi, Alex Deng and Lukas Vermeer worked it out in a 2022 paper written to correct common misreadings of A/B tests. Run 100 tests at the usual settings, 80% power and a two-sided 5% significance threshold, where one idea in ten really works. About eight of the ten real improvements show up as winners. And about two (2.25, on average) of the ninety ideas that did nothing also show up as winners, by chance. So about 22% of the winners are false PUBLISHED .

SHARE OF IDEAS THAT REALLY WORK	SHARE OF WINNERS THAT ARE FALSE, AT 80% POWER
33% (Microsoft)	5.9%
20%	11.1%
15% (Bing)	15.0%
10% (Booking.com, Google, Netflix)	22.0%
8% (Airbnb search)	26.4%

PUBLISHED Kohavi, Deng and Vermeer, “A/B Testing Intuition Busters,” 2022. Counts a winner as a significant result in the right direction.

That’s at 80% power. Cut the power and it gets much worse, because fewer real improvements show up while the chance winners keep coming. Goodson’s 2014 illustration: in 100 tests with 10 real effects, tests stopped at two weeks with under 30% power produce about three real winners and five false ones. “63% of your winning tests are completely imaginary” REPORTED . (Goodson counts 5% of the ideas that did nothing as winners; on the stricter count above it would be about two.) The 2022 paper takes apart a widely shared test that claimed a 337% lift from 82 and 75 visitors. Its power to detect a 10% change was 3%. Even if one idea in three worked, a win from a test like that would be false 63% of the time PUBLISHED .

RUN YOUR NUMBERS

How many of your winners are real?

Example numbers. Replace with yours. Take power from the planner in chapter 3; a test stopped early has less than planned.

Share of your ideas that really work

10 %

Power of your tests

80 %

Significance threshold, two-sided

5 %

8.0	2.3	22%	2.0
real winners, per 100 tests	false winners, per 100 tests	of your winners are false	real improvements missed, per 100 tests

> **About 22% of your winners are false, roughly one in five. Confirm wins that matter before building on them, and be most suspicious of the biggest.**

A false winner here is a test of a change that did nothing, significant in the winning direction by chance: half the threshold. This assumes you don't stop early; peeking adds more false winners on top.

Real winners shrink

Even when a win is real, the lift you measured is usually too big. A small test only reaches significance when chance happens to push the result up, so the winners it produces are the lucky draws. The 2022 paper calls it the winner's curse: "the 'lucky' experimenter who finds an effect in a low power setting" is "cursed by finding an inflated effect" PUBLISHED . The lower the power, the worse it gets: a significant win from a test with 50% power overstates the true lift by about 40% on average, and at 25% power it roughly doubles it DERIVED . Plan on real lifts being smaller than the test said, and treat a surprising one with Twyman's law, which Kohavi quotes often: "Any figure that looks interesting or different is usually wrong" PUBLISHED .

> **The bigger the lift a small test reports, the less you should believe it.**

Three habits that make it worse

01

Peeking

Checking results as they come in and stopping when one side looks ahead. The statistician Evan Miller showed in 2010 that checking after every visitor in a small test of a change that did nothing pushed the false-positive rate from 5% to 26.1% **REPORTED** . A team at Optimizely, which sells a testing tool, simulated millions of tests comparing identical pages and found more than 57% declared a winner or loser at some point during the run **REPORTED** . Johari and colleagues found the error rate can easily rise fivefold even at 10,000 visitors **PUBLISHED** .

02

Many metrics, many segments

Look at 20 metrics or 20 customer segments and, on average, one will show a “significant” difference by chance. Pick the primary metric before launch. Treat segment findings as ideas for the next test, not results.

03

Believing the dashboard’s method fixes it

Some tools use sequential statistics built for checking results as they arrive. Optimizely has since 2015, and VWO added a correction for peeking to its Bayesian engine in December 2024 **REPORTED** . That works, so find out which method your tool uses. But a Bayesian readout isn’t immune to peeking on its own: in David Robinson’s 2015 simulation of a Bayesian stopping rule, checking daily raised false acceptances from 2.5% to 11.8% **REPORTED** . And no method creates power the traffic doesn’t have.

Do this

- **Take your last five winners.** For each, put its traffic and run length into the planner in chapter 3, with the lift you’d have needed to act on (10% if you never set one), not the lift the test reported, to find the power it had. Then put that power into the tool above. Any winner from a test under 50% power goes on a list to confirm, starting with the one you’ve built most on.

BEFORE YOU BELIEVE A RESULT

Check the split, check the size, read the range, then ask whether the story holds. In that order, every time.

A test result is only as good as the machinery that produced it. Two checks decide whether there's a result at all. Two more decide how much to believe it.

1. The split

If you asked for 50/50 and got 52/48 on a big test, something is broken. Aleksander Fabijan and colleagues found this problem, called a sample ratio mismatch, in about 6% of experiments at Microsoft and about 10% of analyses at LinkedIn. It "in most cases completely invalidates experiment results" PUBLISHED , because whatever pushed visitors out of one version usually pushed out a particular kind of visitor. Microsoft flags a mismatch when a standard chi-square test on the counts gives a p-value below 0.0005 PUBLISHED . On a store, the usual causes are a version that redirects to a new page and loses people during the extra load, bot filtering that treats the versions differently, and a script that fails to fire in one version.

2. The size

Did the test reach the size planned before launch, and run whole weeks? If not, it isn't finished, whatever the dashboard says. Never stop early because it looks like a win, unless your tool uses a sequential method designed for it and you know that it does. Stopping early because the split is broken or a guardrail is in trouble is fine.

3. The range

Read the confidence interval, not the single number. A lift of 11% with a range from 1% to 22% says the change probably helps, by an amount you don't know well. Plan on the low end. A range that includes zero says you can't tell, which isn't the same as saying there's no difference.

4. The story

Does the result make sense, and does it hold within the test? Look at it week by week. A lift that's big in week one and gone by week three may be novelty: Kohavi and colleagues describe returning users who "investigate the new feature, click everywhere, and thus introduce a 'novelty' bias that dies quickly" PUBLISHED . Most store visitors are new, so novelty matters less on a store than on a product people use daily, but check anyway. Check guardrails: order value, returns, contribution. And if the result is surprising, run it again. The 2022 paper recommends a stricter threshold, 0.01 or 0.005, before believing surprising results PUBLISHED .

> **Check the split before the result. A broken split means there is no result.**

FROM MY WORK One more check belongs before the test starts: is the baseline clean? On one program, a broken product block was dragging down the numbers of the very thing we planned to redesign. Fixing the bug and testing the redesign at the same time would have credited the redesign with the fix. The rule I hold to: fix first, in both versions, then give the page two to three weeks so you know its new conversion rate before you size the test. Never let a repair ride along in only one version, and never plan a test on a baseline you haven't measured.

RUN YOUR NUMBERS

Read a finished test

Example numbers. Replace with yours. A is the current version, B the change.

Visitors in A

Orders in A

30000	750
Visitors in B	Orders in B
30120	840
Share of traffic meant for A	Planned visitors per version
50 %	30000

OK the split, 49.9% to A (chi-square p 0.625)	2.50 / 2.79% conversion rate, A then B	+11.6% measured lift	1.2% to 22.9% 95% confidence interval for the lift
0.027 p-value, two-sided			

> **B probably helps. The best guess is an 11.6% lift, but anything from 1.2% to 22.9% fits the data, and a significant win usually overstates the lift. Plan on the low end, check order value and returns, and confirm it if you'll build on it.**

Uses a two-proportion test for the p-value, a log-ratio interval for the lift, and a chi-square test on the split; the verdict follows the interval. It assumes you didn't stop early. For tests with more than two versions, compare each with A and use a stricter threshold.

Run an A/A test

Before you trust a testing setup, test it against itself. Split traffic between two identical versions for two weeks. You should see an even split, and usually no winner. A single A/A winner happens one time in twenty by chance, so rerun it. If it wins again, or if the split fails the check, fix the setup before running anything that matters.

Do this

- **Put your last finished test through the reader.** Then schedule an A/A test on your testing tool for the next two weeks, on your highest-traffic page.

PART TWO · WHEN YOU CAN'T TEST · CHAPTER 6

THE LADDER OF EVIDENCE

An A/B test is the top rung, not the only one. Pick the highest rung your traffic can reach, and use it with care.

If the traffic table told you your store can't test most changes, you still have to change things. The choice isn't between a perfect experiment and guessing. There's a ladder of methods between them, each cheaper and weaker than the one above.

01 **An A/B test to its planned size**

Random split, planned size, one primary metric, read with the checks in chapter 5. The strongest evidence, and the most expensive in traffic. Use it for bold changes on your highest-traffic pages.

02 **A holdout rollout**

Ship the change to most visitors and keep a random slice, say 10%, on the old version for a longer period. If the change helps, only the held-back slice misses out while you learn. If it hurts, most visitors bear it, so watch the guardrails weekly. A 90/10 split needs almost three times the visitors of a 50/50 test to reach the same power, so it runs for months, not weeks. Assign the holdout by customer ID where you can, because browser cookies expire over months.

03 **On-off weeks**

Switch the change on and off a whole week at a time, for six to eight weeks, deciding which week of each pair is on by a coin flip rather than strict alternation. It's weaker than randomizing visitors: a sale or a press mention can land in one kind of week, returning shoppers see both versions, and you have only a handful of weeks to compare. Compare week by week rather than pooling visitors, and log everything else that happened. Never use it for prices.

04 **Before and after, with guards**

Compare four whole weeks before the change with four after, with the guards below. The weakest rung that still deserves the name evidence.

05 **Research, then a fix**

For problems that research shows plainly, like a broken button or a missing delivery date, fix them and move on. Chapter 8.

> **A weaker method used carefully beats a stronger one used badly.**

The guards for a before-and-after

A plain before-and-after comparison fools you whenever anything else changed at the same time: the season, the traffic mix, a promotion, a competitor's sale. The guards don't make it an experiment. They make it harder to fool yourself.

- **Write down in advance** the metric, the weeks you'll compare and the size of change you'd act on.
- **Use whole weeks,** and avoid weeks with promotions, launches or holidays on either side.
- **Find a comparison that the change didn't touch.** If you changed the mobile product page, desktop is your comparison: did mobile improve more than desktop over the same weeks? If you changed one product's page, compare with similar products whose pages you left alone.
- **Check the traffic mix.** If paid social's share of traffic doubled in the after period, conversion moved for that reason. Compare channel by channel.
- **Check last year.** If the same weeks rose by a similar amount last year, the season did it.

- **Know your noise.** Run the same comparison on three or four past periods when nothing changed. If the gap between the two moved by 10% in ordinary months, a 9% result tells you nothing.

Say you rebuild the mobile product page. Over four weeks before and four after, mobile conversion goes from 1.6% to 1.8%, a 12.5% lift. Desktop, untouched, went from 2.8% to 2.9% over the same weeks, about 3.6%. Mobile rose about 9% more than desktop (1.125 ÷ 1.036), and that's your best estimate of what the rebuild did. It's weaker than a test, and much better than the 12.5% you'd have reported without the comparison. But check it against noise: at 50,000 mobile visitors a month, a result like this could plausibly be anywhere from a small loss to a 25% gain. That's what the last guard is for.

Test only bold changes

If the page gets about 50,000 visitors or more in four weeks, you can still A/B test, as long as the change is big enough to clear the smallest lift you can see. That rules out button colors and word tweaks and rules in whole new pages, a different offer, a different way of presenting price or delivery. A bold change that loses teaches you more than a timid one you can't read.

Do this

- **For each change on your roadmap,** write which rung you'll use to judge it, and for before-and-after changes, the comparison and the weeks. Changes that can't name a rung are opinions; label them as such in the log.

PART TWO · WHEN YOU CAN'T TEST · CHAPTER 7

RESEARCH THAT FINDS THE LEAK

Five customer sessions, one survey question, your funnel by device and a hard look at three competitors. Most stores find more in a week of this than in a year of tests.

Testing tells you whether a change worked. Research tells you what to change. Small stores need research more than big ones, because when you can only test a few things a year, the quality of your ideas decides everything.

The funnel, by device

Start with the numbers you already have. For the last four weeks, split by phone and desktop: sessions, product page views, add to cart, reached checkout, entered shipping, entered payment, ordered. Find the biggest single drop on each device. That step is where you watch people in the next section.

Watch five people

In 1993 Jakob Nielsen and Thomas Landauer published a model of how many usability problems a test finds as you add participants. Nielsen's popular summary is that five users uncover about 85% of the problems, and that three rounds of five, fixing between rounds, beat one round of fifteen PUBLISHED . It's a model, not a guarantee: when Laura Faulkner drew random groups of five from 60 participants, some groups found 99% of the problems and others 55% PUBLISHED . But the point stands. You don't need many people to find the big problems.

Recruit five people who fit your customer and have never used your site. Give them a phone and a task, like "buy something for your sister's birthday under $50," and ask them to think out loud. Don't help. Watch where they hesitate, what they tap that isn't a link, what they look for and can't find. Record it if they agree. The first session usually finds something the whole team missed.

> **You don't need many people to find the big problems. You need to watch, and not help.**

Ask the people who just bought

A one-question survey on the order confirmation page, or in the first email after purchase, finds the objections that almost cost you the sale. The conversion research firm CXL suggests surveying recent first-time buyers within days of purchase, aiming for at least 100 responses, and asking questions like "What made you almost not buy from us?" REPORTED . Read every answer and group them. The biggest groups are your product page's missing answers.

Four more sources

- **Session recordings and heatmaps:** watch 30 recordings of mobile visitors who reached a product page and left. Look for repeated taps, long pauses and scrolling up and down in search of something.
- **Support tickets and chat logs:** the most common questions before purchase are information your product page should carry.
- **Your 3-star reviews:** buyers who liked the product enough to be fair about it and cared enough to explain what disappointed them.
- **A competitor teardown:** FROM MY WORK I buy from the two or three stores my client's customers compare them with, on a phone, from the ad to the confirmation email, and screenshot every step. Then I list what each competitor answers earlier or more clearly: delivery dates, sizing, returns, reviews, price per serving. The gaps are the list.

A heuristic review, where people who know usability judge your pages against a checklist, is useful too, and Nielsen Norman Group recommends three to five reviewers working separately. It also warns that such reviews "cannot replace user research" REPORTED . Use it to prepare for the five sessions, not instead of them.

What comes out

One ranked list of problems, each with its evidence: how many of the five sessions hit it, how many survey answers mention it, where it sits in the funnel. Problems that are plainly broken

go to chapter 8. Problems with an uncertain fix go to the ladder in chapter 6.

Do this

- **This week, add the post-purchase question** and book five sessions. Next week, run the sessions on a phone, starting on the step where your funnel drops most.

PART TWO · WHEN YOU CAN'T TEST · CHAPTER 8

FIX IT, DON'T TEST IT

Bugs, slowness and hidden information don't need an experiment. Testing them spends your scarcest resource proving what you already know.

Some changes are uncertain: nobody knows whether a new layout will help. Others aren't: a button that doesn't work on some phones is costing you orders, and nobody needs a test to know it. Treating the second kind like the first is one of the most expensive habits in conversion work.

What goes straight on the fix list

- **Errors and bugs.** In Baymard Institute's survey of US shoppers, 17% said they'd abandoned an order because of website errors or crashes REPORTED . Check the checkout on every major phone and browser you get traffic from, every month.

- **Slowness.** Google's Core Web Vitals give the thresholds for a good experience, measured on 75% of page loads: the main content loads within 2.5 seconds, the page responds to a tap within 200 milliseconds, and the layout doesn't jump around, a shift score of 0.1 or less PUBLISHED . Shopify's own help pages name installed apps as one of the biggest factors in store speed, and note that uninstalling an app doesn't automatically remove its code from your theme PUBLISHED .
- **Missing answers.** Anything your five sessions or your survey show people looking for and not finding: the delivery date, the total cost, the returns policy, the sizing. Chapter 9.
- **Unnecessary steps.** Baymard finds the average US checkout shows 23.48 form elements, where 12 to 14 would do, and that the average large site could raise its conversion rate by 35.26% through better checkout design REPORTED . That's an estimate from their benchmarking, not an experiment, but the direction of every item on it is clear: fewer fields, guest checkout, express wallets.

Speed, with the evidence sorted

Speed is where the evidence is most often overstated, so sort it. The strongest example is Vodafone's: a true A/B test in Italy, with half the traffic sent to an optimized landing page. Its main content loaded 31% sooner, by the measure Google calls Largest Contentful Paint, and sales rose 8% REPORTED . The widely quoted Deloitte study for Google, which found that a 0.1-second improvement went with 8.4% more retail conversions, is observational: it watched speed vary naturally across 37 brands REPORTED . And the old line that 53% of mobile visits are abandoned after three seconds comes from 2016 Google data with thin methods; it's too dated and loosely sourced to plan around. Speed matters. Measure your own before and after.

> **If five of five people trip on it, fix it. Save your traffic for questions you can't answer by watching.**

When a fix should still be tested

A few "obvious" fixes have real downsides and deserve a test or a holdout: anything that changes price or what the customer pays for shipping, anything that removes information some shoppers rely on, and anything that trades conversion for margin. Everything else, fix.

And fix before you test anything else on the same page. A test run on top of a broken baseline credits the new version with the repair. Log each fix with its date, so the before-and-after in chapter 6 has something to read.

Do this

- **Make the fix list** from your funnel, your five sessions and a speed check of your top three pages on a phone. Ship the top three fixes this sprint, each logged with the date it went live.

PART THREE · THE PAGE · CHAPTER 9

THE PRODUCT PAGE

Most product pages lose sales by hiding answers. The fixes are rarely clever. They're usually a line of text moved closer to the button.

A stranger on a product page asks the same questions in the same order: what is it, is it for me, can I trust it, what will it cost me in total, when will it arrive, and what happens if I'm wrong. Most pages answer the first at length and bury the rest in tabs, footers and the checkout.

What the benchmark finds

Baymard Institute reviews the product pages of leading ecommerce sites against its usability guidelines. In its benchmark of more than 155 sites, updated in March 2026, 52% of desktop and 62% of mobile product pages rated “mediocre or worse” REPORTED . Some of the gaps it counts:

GAP	SHARE OF SITES
Don’t show the total cost near the buy section	67%
Don’t respond to negative reviews	89%
Size choices not shown as buttons	57%
No link to the returns policy from the product page	44%
No images that show the product’s scale	37%

REPORTED Baymard Institute, “The current state of ecommerce product page UX,” updated March 2026.

Read the list as a set of questions left unanswered. The shopper who can’t see the total cost finds it at checkout, and extra costs are the most common reason US shoppers give for abandoning one: 40% in Baymard’s survey REPORTED . The shopper who can’t find the returns policy wonders what happens if the size is wrong, and leaves to think about it.

Reviews

Northwestern University’s Spiegel Research Center analyzed review data supplied by PowerReviews, a review vendor. At one high-end gift retailer, a product with five reviews was 270% more likely to be bought than one with none, and displaying reviews raised conversion 190% for lower-priced products and 380% for higher-priced ones. Across several product categories, purchase likelihood peaked for average ratings between 4.0 and 4.7 and fell as ratings approached 5.0 REPORTED . The data is observational and much of it comes from one store, so don’t bank the percentages. Take the pattern: reviews matter more as the price rises, and perfection looks suspicious.

> **Show the imperfect reviews and answer them. Perfection reads as a filter.**

The first screen on a phone

On a phone, many visitors never scroll past the first screen, and Contentsquare finds nearly two-thirds of visitors who land on a product page leave without going further REPORTED . What belongs on that first screen, in this order:

01 **The product, in use**

A photo that shows scale and context, not only the product on white.

02 **Who it’s for, in one line**

The problem it solves, in the customer’s words from your reviews and survey.

03 **The review count and average**

Linked to the reviews, including the critical ones.

04 **The price, and the price per use for consumables**

“$38, about $1.27 a serving.” If there’s a subscription option, its price and terms beside it.

05 **The delivery date and the return promise**

“Arrives Thursday. Free returns for 30 days.” Next to the button, not in a tab.

06 **The add-to-cart button, and an express wallet**

Big enough for a thumb.

Everything else, the story, the ingredients, the founder’s note, goes below for the people who scroll. They’re the people who were going to read it anyway.

Fewer choices isn’t always better

Many teams cut variants because of the famous jam study, in which 30% of shoppers who stopped at a display of six jams bought, against 3% at a display of 24 PUBLISHED . A 2010 analysis of 50 experiments on the same question found the average effect was “virtually zero,” with big differences between studies PUBLISHED , and a later review found that too much choice hurts under particular conditions: complex options, unclear preferences, hard decisions PUBLISHED . So don’t cut variants on principle. Watch your five sessions: if people struggle to choose, add a recommendation (“most people start with the 30-day size”) before you remove anything.

Do this

- **Open your top product page on a phone** and check the six items above against the first screen. Anything missing, or more than one scroll away, goes on the fix list in chapter 8.

PART THREE · THE PAGE · CHAPTER 10

THE MOBILE GAP

Phones bring most of the traffic and convert at well under the desktop rate. Some of that gap is how people shop. Some of it is your site.

For most DTC brands, the phone is where the customer meets the store, and it’s where the store performs worst. Closing part of that gap is often the largest conversion opportunity I find in an audit.

The size of the gap

Contentsquare's 2026 benchmark, drawn from 99 billion sessions on more than 6,500 sites in the last quarter of 2025, found phones brought 69.9% of traffic. Desktop converted at 3.4% and mobile at 2.0%; for retail sites, 3.7% against 2%. Mobile sessions lasted 2 minutes 20 seconds, desktop 4 minutes 46 REPORTED . Contentsquare's customers skew toward large sites. For Shopify stores, Littledata's 2023 benchmark of about 2,800 sites put the average conversion rate at 1.4%, with mobile at 1.2% and desktop at 1.9% REPORTED . Both are vendors reporting on their own customers.

Don't aim for parity. Part of the gap is behavior: people browse on their phones in spare minutes and finish on a laptop, and some of those orders get counted as desktop. But part of it is friction, and the benchmarks say there's a lot. Baymard's review of more than 150 mobile sites, updated July 2026, rated 75% of them "mediocre." 93% didn't use adaptive error messages on forms, and 54% had no address lookup or validation REPORTED .

FROM MY WORK A finding I see often in store audits: phones bring most of the visits and convert at a fraction of the desktop rate, while the team reviews the site on a laptop. The plan that follows puts the mobile product page and mobile checkout first, and gives mobile conversion its own line on the weekly report and its own owner.

> **If the team reviews the site on laptops and customers shop on phones, the team is reviewing the wrong site.**

Where the thumb gives up

- **Typing.** Offer express wallets (Apple Pay, Google Pay, Shop Pay, PayPal) on the product page and in the cart. Use the right keyboard for each field and let the browser fill in addresses and cards. Add an address lookup.
- **Errors.** Show what's wrong next to the field, as it happens, in plain words.
- **Tap targets.** Buttons and links big enough for a thumb, with space between them.
- **Pop-ups.** A full-screen pop-up on arrival covers the thing they came to see. Google has discouraged intrusive pop-ups on mobile since January 2017, while allowing banners that use a reasonable amount of screen PUBLISHED . More in chapter 12.
- **Speed.** Measure on a mid-range phone on mobile data, not on the office Wi-Fi.

Measure it properly

Because some phone shoppers finish on a laptop, mobile orders understate what mobile does. Watch the mobile funnel's steps too: add to cart, reached checkout, completed checkout. A fix that raises mobile checkout completion is working even if some of the extra orders show up under desktop.

Do this

- **Once a month, buy from your own store on a phone,** on mobile data, with a card you've never used there, starting from an ad. Time it. Write down every moment you had to type, wait, pinch or hunt.

PART THREE · THE PAGE · CHAPTER 11

SHIPPING, OFFERS AND PRICE TESTS

The changes that move conversion most are the ones that cost margin. Judge them on what they leave, and test prices in a way you'd be comfortable explaining.

Free shipping, a discount, a lower price: each will usually raise conversion. That's what makes them dangerous in a conversion program. A test that

reports conversion rate will declare each of them a winner, and the business can be worse off.

Free shipping

Extra costs are the most common reason US shoppers give for abandoning a checkout: 40% in Baymard's survey named shipping, tax or fees, and 12% said they couldn't see or calculate the total up front REPORTED . So showing the shipping cost early is a fix, not a test.

Making shipping free is a different decision. Michael Lewis, Vishal Singh and Scott Fay studied an online retailer that tried many shipping-fee schedules and found shoppers very sensitive to shipping charges, in whether they ordered and how much. Free shipping and free-shipping thresholds generated extra sales. But the lost shipping revenue, and the segments that didn't respond, made those promotions unprofitable for that retailer PUBLISHED . A 2020 study by Edlira Shehu, Dominik Papies and Scott Neslin found that free-shipping promotions led people to buy riskier products and return more of them, which also made the promotions unprofitable PUBLISHED .

> **A conversion test on a free-shipping offer will almost always say yes. Ask the contribution question instead.**

If you set a free-shipping threshold, put it a little above your typical order, show shoppers in the cart how far they are from it, and judge it on contribution per visitor, returns included. The Whole Machine has a contribution calculator, and The First Offer covers discounts and offers in depth.

Price tests

Random price tests aren't generally illegal in the US, but the rules vary by state and country and are changing. They're also easy to get wrong in ways customers remember. In September 2000, Amazon was found to be showing different customers randomly different discounts, 20% to 40% off, on 68 DVD titles over five days. After the backlash it refunded 6,896 customers an average of $3.10 each, and Jeff Bezos said, "We've never tested and we never will test prices based on customer demographics." Amazon also said any future random price test would give buyers the lowest test price automatically REPORTED .

The law has moved since. New York's Algorithmic Pricing Disclosure Act, in force since November 2025, requires a disclosure when a price is set by an algorithm using the customer's personal data PUBLISHED . Whether a random price test that uses no personal data falls under it hasn't been settled; ask counsel if you sell into New York. At the federal level, the FTC published initial findings in January 2025 from a study of how companies use personal data to set individual prices, without declaring the practice illegal PUBLISHED . One Shopify price-testing app, Intelligems, sets the store's list price to the highest price in the test and applies test prices at checkout, so ads and shopping feeds never show a lower price than the store REPORTED .

Four rules keep a price test defensible:

01 **Randomize; never target**

Assign prices at random, not by device, location or anything you know about the person.

02 **One price per person**

A visitor who comes back sees the same price.

03 **Honor the lowest**

Be ready to refund the difference to anyone who paid more than the lowest tested price, and consider doing it automatically when the test ends, as Amazon promised in 2000.

04 **Judge on contribution per visitor**

Not conversion rate. A higher price can lose orders and still win.

And be realistic about size. A price test is a revenue test, and chapter 3 showed that revenue needs several times the traffic of conversion. Most stores should test prices in bold steps, across a product line at once so the test has enough traffic, and only with visitor-level randomization that keeps each person on one price. Don't use on-off weeks for price: returning shoppers would see the price change from week to week.

Do this

- **List every offer and shipping change of the last year** and ask what it did to contribution per visitor, not conversion rate. If nobody can say, re-read the two you're most confident about with the before-and-after guards.

PART THREE · THE PAGE · CHAPTER 12

QUIZZES, POP-UPS AND SEARCH

Three features sold with big conversion numbers. Each can help. Most of the headline numbers don't measure what the feature caused.

Quizzes, pop-ups and site search share a statistical trap. The people who use them were already different from the people who don't. Compare users with non-users and the feature gets credit for the difference.

Quizzes

Quiz vendors report that quiz takers convert several times better than other visitors. The quiz vendor Octane AI's site claims "4x higher conversion," with no sample, period or comparison group disclosed REPORTED . Shoppers who take a quiz are engaged by definition. The only way to learn what a quiz adds is to show its entry points to a random half of visitors and compare the halves.

A quiz's lasting value is often the answers, not the conversion. Store each shopper's current answers on their profile in your email platform, so segments and emails can use them, and store each completed quiz as a dated event, so you can see answers change and trigger flows from them. Version the quiz, so a changed question doesn't silently break every segment built on it. And drop any question you can't name a use for. Every extra question costs completions. I've written the full data model up as a note: Most brands that have quiz data cannot build one segment from it.

Pop-ups

Omnisend, an email platform that sells pop-ups, reports that across 1.24 billion pop-up views in 2025, the average sign-up rate was 2.1%: 2.0% for a standard form, 2.3% for a multi-step one and 3.5% for a spin-to-win wheel REPORTED . A pop-up grows the list, which is worth a lot. It can also cost orders by covering the product on arrival, and Google discourages full-screen pop-ups on mobile (chapter 10).

So on phones, prefer a small banner, or a pop-up that waits until a second page view or real scrolling. Judge the pop-up on both sides of the trade: sign-ups gained, and conversion on the pages where it shows, against a random slice of visitors who never see it.

> **Users of a feature were different before they used it. Compare random halves, not users with non-users.**

Search

"Shoppers who search convert two to three times as often" traces to blog posts from 2013 and 2015 that don't say how it was measured REPORTED , and it measures intent: people who search already know what they want. The better reason to care about search is that it's often broken. Baymard's 2026 benchmark of more than 170 sites found mediocre-or-worse search on 46% of desktop and 58% of mobile sites. 66% failed to handle searches for things other than products, like "returns" or "shipping," and 54% failed on abbreviations or symbols REPORTED .

The fix is a weekly habit. Pull the searches that returned no results, add synonyms and redirects for the common ones, and make sure the search box is visible on mobile, not hidden behind a menu.

Do this

- **For whichever of the three you use,** find the number you've been judging it by. If it compares users with non-users, replace it with a random holdout at the next chance.

PART FOUR · RUNNING A PROGRAM · CHAPTER 13

WHAT ONE GOOD TEST LOOKS LIKE

The most famous website test in politics measured a 40.6% lift. It also produced a $60 million headline that nobody measured. Both halves are the lesson.

In December 2007, Dan Siroker, then working on Barack Obama's presidential campaign and later a co-founder of the testing company Optimizely, ran a test on the campaign's splash page, the page that asked visitors to sign up with their email address before entering the site. He wrote it up in 2010 on Optimizely's blog, and the story has been told in marketing talks ever since.

What was tested

Two parts of the page changed at once. Four versions of the button text ("Sign Up," "Learn More," "Join Us Now" and "Sign Up Now") were crossed with six pieces of media: three

photos and three videos. That's 24 combinations, shown to 310,382 visitors. The winner, a "Learn More" button with a family photo, raised the sign-up rate from 8.26% to 11.6%, a lift of 40.6%. And every video did worse than every photo REPORTED .

Why it was a good test

- **The traffic fit the question.** Split 24 ways, each combination saw about 12,900 visitors and about 1,070 sign-ups, enough to see a difference of about 12% between two combinations, or about 16% once you allow for comparing 23 combinations with the original DERIVED . The winning lift was more than twice that.
- **The changes were bold.** Not shades of a button, but different media and different asks. Bold changes are the only kind a limited amount of traffic can read.
- **It measured the outcome that mattered.** Sign-ups, the thing the page existed for, not clicks on a video or time on page.
- **It was on the busiest page.** Most visitors met the splash page first, so the test borrowed most of the site's traffic.

What it didn't measure

The post's title says the test raised $60 million. It didn't measure that. Siroker estimated that about 10 million people signed up over the campaign, assumed the lift held the whole time, and calculated 2.88 million extra email addresses. He then assumed 10% of sign-ups volunteered, for 288,000 extra volunteers, and that each address was worth an average of $21 in donations, for $60 million REPORTED . Each step is reasonable. None was measured by the test, and the post reports no significance test or range for the lift itself. With 24 combinations, the best one also benefits from the winner's curse in chapter 4: picking the top of 24 flatters it.

> **Keep the measurement and the extrapolation in separate sentences. The first is evidence. The second is a forecast.**

That's not a criticism of the test, which was well designed and clearly worth running. It's a lesson in how results travel. A measured 40.6% lift on sign-ups became "$60 million" in the headline, and the headline is what people remember. In your own reports, write

the measured number first, with its range, and put the projection after it with each assumption named.

Do this

- **Find your splash page:** the page most of your visitors pass through. Usually it's the home page or your best-selling product page. Plan the one bold test it can carry this quarter with the planner in chapter 3, and write the report's first sentence before the test starts: "We measured ___ on ___, with a range of ___."

PART FOUR · RUNNING A PROGRAM · CHAPTER 14

THE BRIEF AND THE LOG

A page of paperwork before each test and a line after it. That's the difference between a team that tests and a team that learns.

A test that won without a written reason teaches you to repeat a result you don't understand. A test that lost with one teaches you something about your customers you'll use for years. The brief is where the reason goes.

The brief

Fill it in before anything launches. The Whole Machine has a shorter version; this one adds the decisions that part one of this guide showed are easy to get wrong.

01 **Because we saw**

The evidence from research: sessions, survey answers, the funnel. “Three of five people couldn’t find the delivery date.”

02 **We believe**

The change, and who sees it.

03 **Primary metric**

One. Usually orders per visitor, or contribution per visitor for anything touching price or shipping.

04 **Guardrails**

What mustn’t get worse: order value, returns, contribution, page speed.

05 **Smallest lift worth acting on**

And the planner’s answer: visitors per version, weeks, and the power you’ll have.

06 **The rung**

A/B test, holdout, on-off weeks or before-and-after, from chapter 6.

07 **Decided in advance**

What you’ll do if it wins, if it loses, and if it can’t tell. Writing the third answer down is what stops a team from rerunning the same test until it wins.

> **Decide what you’ll do with every possible result before you see any of them.**

The log

One row per test and per major site change, in a shared sheet anyone can search: the date, the page, the brief, the rung, the planned and actual size, the result with its range, and what changed because of it. Before anyone proposes a test, they search the log. A year in, the log is the most valuable document in your conversion program, because it’s the only record of what your changes did, as opposed to what someone remembers they did.

How many tests

A page that needs four weeks per test can carry at most thirteen tests a year, one after another, and fewer once you allow for launches and sales. Tests on different pages can run at the same time, as long as each assigns visitors at random on its own and the two changes can't clash (two tests both changing the delivery message, say). Counted against the traffic table, most DTC stores I see have room for somewhere between two and twenty real tests a year. Spend them on bold changes drawn from research, not on small ideas drawn from opinion.

FROM MY WORK For experiences a testing tool doesn't reach, like email flows and post-purchase pages, I stamp every customer with a random digit from 0 to 9 when they're first recorded, and never change it. Holding out digit 0 gives a standing 10% control group, with no setup per test. It measures everything you've shipped to the other 90% together, so read it as the combined effect of your program. To read one change on its own, give it its own random split.

Do this

- **Start the log today** with your last five tests and your last five major site changes. Fill in what you can. The blanks show you which parts of the brief your team has been skipping.

PART FOUR · RUNNING A PROGRAM · CHAPTER 15

THE FIRST THIRTY DAYS

Numbers first, then customers, then fixes, then one test worth running. Four weeks, in that order.

Whether you're starting conversion work or restarting it, the order is the same: learn what your traffic can support, learn what's wrong from the people who shop, fix what's broken, and only then spend traffic on a test.

01 **Week one: the numbers**

Four weeks of visitors, conversion and orders for your top three pages, by device, placed in the traffic table (front). The funnel by device (chapter 7). Buy from your own store on a phone and time it. Add the post-purchase question. If you have a testing tool, start a two-week A/A test on your busiest page (chapter 5).

02 **Week two: the customers**

Five sessions on a phone, starting at your biggest funnel drop. A competitor teardown. Read the pre-purchase questions in support tickets and your 3-star reviews. End the week with one ranked problem list.

03 **Week three: the fixes**

Ship the top fixes from the list, each logged with its date (chapter 8). The pages you fixed now need two to three weeks to settle into a new baseline.

04 **Week four: one test and the log**

Pick one bold change for a page that can carry it and that you didn't fix this month, or wait until a fixed page has two weeks of new baseline. Run the planner, write the brief (chapter 14) and launch once the A/A test is clean. Start the log. Re-score the audit.

At day thirty you'll have fewer tests running than most teams, and more reason to believe the ones you have. The fixes will already be live and logged. And the ranked problem list will keep you busy for a quarter.

> **Research first, fixes second, tests third. Most teams run it backwards.**

Do this

- **Put the four weeks on the calendar today,** with an owner for each line, and book the five sessions before you close this page.

CLOSE

DAY ONE

Six things the person who owns conversion needs on the first day.

Whoever owns conversion, a new hire, an agency, or you on the Monday you decide to stop guessing, needs six things on day one. Without them, the first month goes on hunting for things they should have been handed.

01 **Analytics access**

With page-level traffic, conversion and a funnel by device, and someone who can explain how it's configured.

02 **The testing tool, if there is one**

With every past test, its settings and its results.

03 **A list of site changes**

Every redesign, app install, theme change, price change and promotion of the last year, with dates. Gaps are findings.

04 **Customer words**

Support tickets, chat logs, reviews and any survey answers, unfiltered.

05 **Contribution per order**

From finance, so offer and shipping changes can be judged on what they leave.

06 **Authority to fix**

A developer's time and the right to ship a fix without a test when research shows it's broken.

Do this

- **Ask for all six in one message** today. Whatever comes back slowest is where the store knows least about itself.

CLOSE

THE SHELF

The books and papers this guide leans on, and what to take from each.

Trustworthy Online Controlled Experiments

Ron Kohavi, Diane Tang and Ya Xu, 2020

The manual. Everything in part one is here in more depth, from people who ran testing at Microsoft, Google and LinkedIn.

Experimentation Works

Stefan Thomke, 2020

How companies like Booking.com built a culture of testing, and what it asks of leaders.

Rocket Surgery Made Easy

Steve Krug, 2009

A short, practical guide to running your own usability sessions, the method in chapter 7.

Don’t Make Me Think, Revisited

Steve Krug, third edition, 2014

The classic on why people get stuck on websites, and still the fastest way to see your own site as a stranger does.

Uncontrolled

Jim Manzi, 2012

On why controlled experiments beat intuition in business, and how rarely ideas work.

Statistical Rules of Thumb

Gerald van Belle, second edition 2008

Where the sample-size rule in chapter 3 lives, with dozens more like it.

And the research: Kohavi, Deng and Vermeer (2022) on false winners; Fabijan and colleagues (2019) on broken splits; Johari and colleagues (2022) on peeking; Deng, Xu, Kohavi and Walker (2013) on variance reduction; Nielsen and Landauer (1993) and Faulkner (2003) on how many users to test; Lewis, Singh and Fay (2006) and Shehu, Papies and

Neslin (2020) on free shipping; Iyengar and Lepper (2000), Scheibehenne and colleagues (2010) and Chernev and colleagues (2015) on choice. Full references are in Appendix C.

CLOSE

ABOUT THE AUTHOR

Andrew Lauchner is a growth and retention operator for consumer brands and the author of sixteen free books on the method. He fixes the second-purchase problem through his practice, Growth Legend, as one engagement: The Second-Purchase Sprint. He wrote The Second Order, on turning first-time buyers into second-time buyers; Close the Loop, on getting customers to bring the next customer; The First Offer, on the offer that wins the first order; The Whole Machine, on the fundamentals of DTC growth; and The Standing Order, on subscription programs.

As Senior Director of Growth and Retention Marketing at Gallery Furniture, he rebuilt the customer journey and the sales playbooks together. He has worked on growth and retention at Binance and 3Commas, and has been Head of Growth and Retention at Greatness Wins and at Nexus Agriscience.

The methods marked “from my work” come from auditing client stores and running their programs. Clients aren’t named and their numbers aren’t here.

WHAT COLLEAGUES SAY

> “Andrew led retention, lifecycle, and email/SMS, but what separates him from most in this space is how deeply he understands the role retention plays in the overall growth engine.”
>
> Akram Khan, Head of Marketing at Gallery Furniture, senior to Andrew but didn’t manage Andrew directly

Andrew answers every note from people trying to make their stores convert. Write to andrew@growthlegend.com or message him on LinkedIn.

APPENDIX A

FOR YOUR ANALYST

The formulas behind the three calculators, and three queries every store should be able to run.

The formulas

FOR	FORMULA	NOTES
Visitors per version	$n = 16 \times p(1 - p) / d^2$	p: current conversion rate. d: absolute change to detect. 95% confidence, 80% power; use 21 for 90% power.
Smallest detectable lift	$L = \sqrt{(16(1 - p) / (p \times n))}$	Relative lift, for n visitors per version.

FOR	FORMULA	NOTES
Power at n	$\Phi(d \times \sqrt{(n / (2p(1 - p))) - 1.96})$	n is visitors per version. Φ is the standard normal cumulative distribution function.
False winners	$(1 - s)(\alpha/2) / ((1 - s)(\alpha/2) + s \times power)$	s: share of ideas that work. α: two-sided threshold. Kohavi, Deng and Vermeer, 2022.
Split check	$\chi^2 = \Sigma (observed - expected)^2 / expected$	One degree of freedom for two versions. Flag $p < 0.0005$.
Two-version result	$z = (p_B - p_A) / \sqrt{\bar{p}(1 - \bar{p})(1/n_A + 1/n_B)}$	$\bar{p}$: pooled conversion rate, for the p-value. The range is for the ratio p_B/p_A: $\exp(\ln(p_B/p_A) \pm 1.96\sqrt{((1 - p_A)/x_A + (1 - p_B)/x_B)}) - 1$, where x is orders. The verdict follows the range.

Randomize and count at the same level. If the tool assigns visitors, analyze visitors, not sessions or page views. A visitor who comes back five times is one visitor, and counting their sessions separately makes the result look more certain than it is.

The funnel by device

```
-- sessions reaching each step, last 28 days, by device
-- events: one row per event, with session_id, device, name, ts
SELECT device,
       COUNT(DISTINCT session_id)                                          AS sessions,
       COUNT(DISTINCT session_id) FILTER (WHERE name = 'view_item')        AS viewed_produ
       COUNT(DISTINCT session_id) FILTER (WHERE name = 'add_to_cart')      AS added_to_car
       COUNT(DISTINCT session_id) FILTER (WHERE name = 'begin_checkout')   AS began_checkou
       COUNT(DISTINCT session_id) FILTER (WHERE name = 'add_shipping_info') AS entered_sh
       COUNT(DISTINCT session_id) FILTER (WHERE name = 'add_payment_info') AS entered_payr
       COUNT(DISTINCT session_id) FILTER (WHERE name = 'purchase')         AS purchased
FROM events
WHERE ts >= CURRENT_DATE - INTERVAL '28 days'
GROUP BY device;
```

The event names follow Google Analytics 4's recommended ecommerce events; rename them to match your setup. The syntax is Postgres; in BigQuery, write `COUNT(DISTINCT IF(name = 'add_to_cart', session_id, NULL))`. Divide each column by the one before to get each step's pass rate, and compare phones with desktop step by step. Express wallets can skip the shipping and payment events, so a step above 100% means an event is missing, not a miracle.

The split check

```
-- visitors assigned to each version, counted once each
SELECT variant, COUNT(DISTINCT visitor_id) AS visitors
FROM exposures
WHERE experiment_id = 'pdp-delivery-date'
GROUP BY variant;
```

Put the two counts into the reader in chapter 5. Run the check daily in the first days of any test, not only at the end: a broken split caught on day two costs two days.

Page traffic for the traffic table

```
-- visitors, orders and conversion for visitors who saw a page type
SELECT page_type,
       COUNT(DISTINCT s.visitor_id) AS visitors,
       COUNT(DISTINCT o.order_id)   AS orders,
       COUNT(DISTINCT o.order_id)::numeric
         / COUNT(DISTINCT s.visitor_id) AS orders_per_visitor
FROM page_views s
LEFT JOIN orders o
  ON o.visitor_id = s.visitor_id
 AND o.ordered_at BETWEEN s.viewed_at AND s.viewed_at + INTERVAL '1 day'
WHERE s.viewed_at >= CURRENT_DATE - INTERVAL '28 days'
GROUP BY page_type
ORDER BY visitors DESC;
```

This counts an order toward a page if it came within a day of the visit. That's an approximation, and a generous one, but it's the right shape for deciding which pages can carry a test.

APPENDIX B

TEMPLATES

Five one-page forms. Copy them into whatever your team already uses.

The test brief

```
BECAUSE WE SAW     evidence from research (sessions, survey, funnel):
WE BELIEVE         the change, and who sees it:
PRIMARY METRIC     one:
GUARDRAILS         order value / returns / contribution / speed:
SMALLEST LIFT      worth acting on:        %
PLANNED SIZE       visitors per version:        weeks:        power:
RUNG               A/B / holdout / on-off weeks / before-and-after
IF IT WINS         we will:
IF IT LOSES        we will:
IF IT CAN'T TELL   we will:
OWNER              name, and the date it will be read:
```

The session script

```
BEFORE    "We're testing the site, not you. Please think out loud.
           There are no wrong answers, and I won't be able to help."
TASK 1    "You're shopping for [need]. Find something you'd buy."
TASK 2    "Buy it. Stop when you'd have to pay."
TASK 3    "Before buying, find out when it would arrive and how
```

```
           returns work."
AFTER      "What was the most confusing moment?"
           "What almost made you leave?"
NOTE       Every hesitation over five seconds, every tap on
           something that isn't a link, every question asked aloud.
```

The post-purchase question

```
ON THE CONFIRMATION PAGE, OR THE FIRST EMAIL AFTER PURCHASE
One question, free text:
  "What almost stopped you from buying today?"
Optional second question, multiple choice:
  "Where did you first hear about us?"
Read every answer weekly. Group them. Count the groups.
```

The test log

```
DATE | PAGE | CHANGE | RUNG | BRIEF LINK | PLANNED SIZE
     | ACTUAL SIZE | SPLIT CHECK | RESULT (WITH RANGE)
     | DECISION | WHAT WE LEARNED | OWNER
```

The before-and-after record

```
CHANGE              what, where, live date:
METRIC              one, decided before the change:
BEFORE WEEKS        whole weeks, no promotions:
AFTER WEEKS         same length, no promotions:
COMPARISON          what the change didn't touch (other device, similar
                    products):
TRAFFIC MIX CHECK   share by channel, before vs after:
LAST YEAR           same weeks' change last year:
RESULT              change relative to the comparison (ratio), and the
                    same figure for past periods with no change:
```

APPENDIX C

SOURCES

Every external source, by chapter. Web sources were read in September 2026.

Experimentation and statistics (chapters 2 to 5, 13, 14)

- Ron Kohavi, Diane Tang and Ya Xu. Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press, 2020. Chapter 1 excerpt.
- Ron Kohavi, Alex Deng and Lukas Vermeer. "A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments." KDD 2022. Full text.
- Ron Kohavi, Thomas Crook and Roger Longbotham. "Online Experimentation at Microsoft." Third Workshop on Data Mining Case Studies, 2009.
- Ron Kohavi, Roger Longbotham, Dan Sommerfield and Randal M. Henne. "Controlled experiments on the web: survey and practical guide." Data Mining and Knowledge Discovery 18(1), 2009.
- Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker and Ya Xu. "Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained." KDD 2012.
- Ron Kohavi, Alex Deng, Roger Longbotham and Ya Xu. "Seven Rules of Thumb for Web Site Experimenters." KDD 2014.
- Ron Kohavi and Stefan Thomke. "The Surprising Power of Online Experiments." Harvard Business Review, September–October 2017.
- Stefan Thomke. "Building a Culture of Experimentation." Harvard Business Review, March–April 2020.
- Robert Lehr. "Sixteen S-squared over D-squared: a relation for crude sample size estimates." Statistics in Medicine 11(8), 1992. Gerald van Belle, Statistical Rules of Thumb, second edition, Wiley, 2008.
- Aleksander Fabijan, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer and Pavel Dmitriev. "Diagnosing Sample Ratio Mismatch in Online Controlled Experiments." KDD 2019. Microsoft Research, "Diagnosing Sample Ratio Mismatch in A/B Testing," 2020.

- Ramesh Johari, Pete Koomen, Leonid Pekelis and David Walsh. “Always Valid Inference: Continuous Monitoring of A/B Tests.” Operations Research 70(3), 2022.
- Evan Miller. “How Not To Run an A/B Test.” 2010.
- Martin Goodson. “Most Winning A/B Test Results Are Illusory.” Qubit, 2014.
- Leonid Pekelis. “Statistics for the Internet Age: The Story Behind Optimizely’s New Stats Engine.” Optimizely, January 2015. VWO, how VWO calculates a winning variation and Enhanced SmartStats, December 2024.
- David Robinson. “Is Bayesian A/B Testing Immune to Peeking? Not Exactly.” 2015.
- Alex Deng, Ya Xu, Ron Kohavi and Toby Walker. “Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data.” WSDM 2013.
- Randall A. Lewis and Justin M. Rao. “The Unfavorable Economics of Measuring the Returns to Advertising.” Quarterly Journal of Economics 130(4), 2015.
- Dan Siroker. “How Obama Raised $60 Million by Running a Simple Experiment.” 2010, republished by Optimizely.

Research methods (chapters 6, 7)

- Jakob Nielsen and Thomas K. Landauer. “A mathematical model of the finding of usability problems.” INTERCHI 1993. Jakob Nielsen, “Why You Only Need to Test with 5 Users,” 2000.
- Laura Faulkner. “Beyond the five-user assumption.” Behavior Research Methods, Instruments, and Computers 35(3), 2003.
- Nielsen Norman Group. How to conduct a heuristic evaluation; A/B testing 101, 2024.
- CXL. On qualitative customer surveys.

The page, mobile and speed (chapters 8 to 10, 12)

- Baymard Institute. Cart abandonment rate and reasons, updated 2025; current state of checkout UX, 2025; current state of product page UX, March 2026; mobile UX, July 2026; ecommerce search, 2026.
- Contentsquare. 2026 Digital Experience Benchmark. Littledata, Shopify conversion rate benchmark, 2023.
- Tom Collinger, Edward Malthouse, Ewa Maslowska and Yorgos Askalidis. “How Online Reviews Influence Sales.” Spiegel Research Center, Northwestern University, 2017.
- Sheena Iyengar and Mark Lepper. “When Choice Is Demotivating.” Journal of Personality and Social Psychology 79(6), 2000. Benjamin Scheibehenne, Rainer Greifeneder and Peter Todd, “Can There Ever Be Too Many Options?” Journal of Consumer Research 37(3), 2010. Alexander Chernev, Ulf Böckenholt and Joseph Goodman, “Choice overload: A conceptual review and meta-analysis.” Journal of Consumer Psychology 25(2), 2015.
- Google. Core Web Vitals; INP becomes a Core Web Vital, 2024; Vodafone case study, 2021; Milliseconds Make Millions, 2020; avoid intrusive interstitials.

- Shopify Help Center. Improving web performance.
- Omnisend. Pop-up statistics, February 2026. Octane AI, site claims, read September 2026. Econsultancy on site search, 2013; Adobe on site search, 2015.

Shipping and price (chapter 11)

- Michael Lewis, Vishal Singh and Scott Fay. “An Empirical Study of the Impact of Nonlinear Shipping and Handling Fees on Purchase Incidence and Expenditure Decisions.” Marketing Science 25(1), 2006.
- Edlira Shehu, Dominik Papies and Scott A. Neslin. “Free Shipping Promotions and Product Returns.” Journal of Marketing Research 57(4), 2020.
- Amazon. Statement regarding random price testing, September 27, 2000.
- New York General Business Law § 349-a. Federal Trade Commission, surveillance pricing study, January 2025.
- Intelligems. Price testing documentation, read September 2026.

THE HONEST TEST

Andrew Lauchner · September 2026 · The Whole Machine · The Second Order · Close the Loop · The First Offer · The Standing Order · Free tools · andrewlauchner.com

Published figures come from peer-reviewed research, books by the researchers themselves, official documentation and law, as of the dates given. Reported figures come from named companies, research firms and outlets; testing and analytics vendors’ benchmarks are marked as vendor data. Derived figures are my arithmetic with a standard formula, inputs shown. Methods from my work come from client stores, with clients unnamed and their figures omitted. Examples that open with Say or Picture use made-up round numbers. Pricing, privacy and consumer-protection rules vary by state and country and change often; have counsel review price tests. None of this is legal or financial advice.

Privacy