

The Conversion Experiment Playbook for Service-Business Sites

A method for running honest A/B tests when traffic is limited: write the hypothesis, fix the metric in advance, be realistic about sample size, set stop rules, and log every result including the flat ones.

Contents

01 Write the hypothesis before you build the variant

02 Sample size on low traffic: do the arithmetic first

03 What to test first on a service-business site

04 Fix the metric, the stop rule and the rule against peeking

05 Run the test without breaking your search visibility

06 Experiment checklist and the results log

Key Takeaways & Sources

01

Write the hypothesis before you build the variant

A test without a hypothesis is a coin toss that you pay to run. The format that keeps teams honest has four parts: the change, the reason, the metric and the expected direction. For example: "Because visitors from search do not see the price range before the form, adding a range above the form will increase completed enquiries." The reason is the important part. It names the belief that is being tested, so a failed result teaches you something about visitors instead of only about a layout.

Good reasons come from evidence you already have rather than from taste. Session recordings, form analytics that show where people leave, answers to a question in an enquiry email, sales call notes and customer interviews are all legitimate sources. Be careful with analytics numbers on their own, because they show where people drop out and not why. Write down the source of the belief next to the hypothesis so a reviewer can see whether it rests on evidence or on a guess.

Keep each test to one change that can be explained in a sentence. If you change the headline, the form and the layout together and the page wins, you will not know which change did it, and you will not be able to reuse the learning on another page.

Good reasons come from evidence you already have rather than from taste.

02

Sample size on low traffic: do the arithmetic first

The number of visitors you need depends on your current conversion rate, the smallest improvement you care about detecting, and how much risk of a false result you accept. Small lifts on low baseline rates need large samples. As an illustration only, with a baseline conversion rate of 3 percent, detecting a relative lift of 20 percent (from 3.0 to 3.6 percent) at conventional settings of 80 percent power and 5 percent significance needs very roughly 14,000 visitors in each variant. Your figures will differ, so use a sample size calculator such as Evan Miller's with your own baseline rather than relying on this example.

For many service-business sites that number is out of reach in a reasonable time. A site with a few thousand visits a month would need to run for most of a year to answer that question for a single page, and a long test is exposed to seasonality, campaign changes and site edits that contaminate the result. This is not a reason to give up on evidence. It is a reason to change the question. Test larger changes, where a bigger effect is plausible and the required sample is smaller. Test on the pages with the most traffic. Use a higher-volume metric earlier in the funnel, such as clicks on the contact button, as a secondary signal while holding enquiries as the primary one. Or accept that a formal A/B test is the wrong tool and use before-and-after comparisons, user testing with five to ten people, and qualitative review, labelled honestly as lower-confidence evidence.

Also be clear about what a result means when the test is underpowered. A flat result from a test that could only have detected a large effect does not show the change did nothing. It shows that any effect was smaller than the test could see. Write that sentence in the log instead of "no difference".

Situation	Realistic approach	How to label the result
Several thousand visits per page per month, ample enquiries	Standard A/B test on a bold change with a calculated sample size	Confirmatory if the plan was followed
Low traffic, enquiries a few per week	Test large changes only; run a fixed duration of full weeks; use a secondary click metric alongside	Directional; state the smallest effect the test could detect
Very low traffic	Skip A/B; use moderated user tests and before-and-after comparison	Qualitative or observational evidence, not a test win
Paid traffic you can scale	Split the landing page on the ad platform with equal budgets and a fixed end date	Confirmatory for that traffic source only

03

What to test first on a service-business site

Start with what the page says and asks, not how it looks. The highest-value early tests are usually about clarity and friction: whether the headline states what the service is and who it is for, whether the price or a range is shown, what the visitor must provide to enquire, and what the page tells them will happen next. These are changes that can alter a visitor's decision, so a real effect is more plausible than from a recoloured button.

Rank candidate tests by three things: how much evidence suggests the problem is real, how much traffic reaches the page, and how cheap the change is to build. A short, honest scoring session beats arguing about favourites. Do not test pages that few people see, and do not test a form that has not been checked for errors on mobile. Fix obvious defects first; broken forms and slow pages do not need a hypothesis, they need a repair.

Be wary of imported lists of "winning" tests from other companies. Their audiences, offers and traffic sources differ from yours, and published case studies can skew toward wins over losses. Treat them as a source of ideas to test, not of expected results.

- Headline and opening paragraph: does a stranger know what you do and for whom?
- Form length and required fields: what is the minimum needed to reply?
- Price transparency: a range, a starting point, or a description of what affects price
- Next-step copy: what happens after the enquiry and how fast
- Proof placement: real, checkable evidence near the point of decision

Rank candidate tests by three things: how much evidence suggests the problem is real, how much traffic reaches the page, and how cheap the change is to build.

04

Fix the metric, the stop rule and the rule against peeking

Choose one primary metric before the test begins and write it down. For a service business it is normally a qualified enquiry, a completed form or a booked call, not page views or clicks. You can track secondary metrics, but if you wait until the data arrives and pick whichever metric looks best, you are guaranteed to find something. Decide too which visitors are included, how bots and internal traffic are excluded, and what counts as a conversion, so the definition cannot drift. Pre-registering a plan in this way is a practice borrowed from research, and the Center for Open Science describes the idea of recording the plan before seeing the data.

Peeking is the quiet failure. If you check a running test every day and stop as soon as a result crosses the significance threshold, the chance of a false positive rises well above the nominal 5 percent, because you gave random noise many chances to look like a win. Evan Miller's article on how not to run an A/B test explains the problem and the arithmetic. The fix is procedural: calculate the sample size or choose a fixed duration, run the test to that point, and do not make a decision before then. If you must monitor for breakage, check that traffic is splitting as expected and that conversions are being recorded, and look at nothing else. Sequential testing methods that allow valid early stopping exist, but they need to be designed into the plan from the start and are not a licence to stop when you like the number.

Write the stop rule in plain terms: "Run for four full weeks or until each variant has 1,500 visitors, whichever is later, then evaluate the primary metric once." Include the circumstances that void a test, such as a site outage, a major campaign change or a tracking fault, and what you will do in each case. Run tests for whole weeks so day-of-week patterns are balanced.

Evan Miller's article on how not to run an A/B test explains the problem and the arithmetic.

05

Run the test without breaking your search visibility

If you split traffic between URLs, take care that search engines are not confused. Google's guidance on website testing says not to cloak, meaning not to show Googlebot different content from users. For tests on separate URLs, it recommends using `rel="canonical"` on the variants pointing to the original, using 302 redirects instead of 301s for temporary redirects, and running the test only as long as needed, then removing the test code and pages promptly.

Keep the experience honest for visitors as well. Do not fake scarcity or invent testimonials as a variant, since a win built on a claim you cannot support will damage trust and, in some cases, breach advertising rules. Before launching, load both variants on a phone and desktop, submit a test enquiry through each and confirm it arrives in the CRM tagged with the variant. If the variants are not both recorded correctly, you are measuring noise.

06

Experiment checklist and the results log

The sequence below is the minimum for a test whose result you will be willing to act on. The accompanying worksheet holds the plan and the log in one place, so the hypothesis, the metric and the stop rule are written before any data exists and cannot be quietly edited afterwards.

Step	Owner	Gate to pass before moving on
Write the hypothesis with change, reason, metric and direction	Test owner	Reason cites evidence already held, such as recordings, enquiry notes or form analytics
Calculate the required sample for the smallest effect worth detecting	Analyst	Duration at current traffic is acceptable, or the test is redesigned or dropped
Record the primary metric, inclusions and exclusions and the stop rule	Test owner	Plan saved with a date before launch; a second person has read it
Build variants and run quality checks on mobile and desktop	Web developer	Test enquiries from each variant arrive in the CRM correctly labelled
Launch and check only for breakage during the run	Analyst	Traffic split is as expected; no outcome numbers are discussed until the stop point
Evaluate once at the stop point against the primary metric	Analyst	Result written as effect with uncertainty, and any secondary findings marked exploratory

Step	Owner	Gate to pass before moving on
Decide: ship, revert or retest, and log the learning	Test owner	Log entry complete, including flat and losing results and the smallest effect detectable

Key Takeaways

- Most service-business sites do not have the traffic to detect small lifts. Plan for bigger changes, longer tests, or a different method.
- Write the hypothesis, the primary metric and the stop rule before the test starts, and do not change them afterwards.
- Looking at results daily and stopping when they look good inflates false positives. Decide the sample size or duration first and wait for it.
- Test the things that change what a visitor understands or is asked to do before testing button colours.
- Log every test, including flat and losing ones. A log of what did not work is what stops the team repeating it.

Sources

1. Website testing and Google Search (A/B testing guidance) — <https://developers.google.com/search/docs/crawling-indexing/website-testing>
2. How Not To Run an A/B Test (Evan Miller) — <https://www.evanmiller.org/how-not-to-run-an-ab-test.html>
3. Sample size calculator for A/B tests (Evan Miller) — <https://www.evanmiller.org/ab-testing/sample-size.html>
4. Preregistration (Center for Open Science) — <https://www.cos.io/initiatives/prereg>