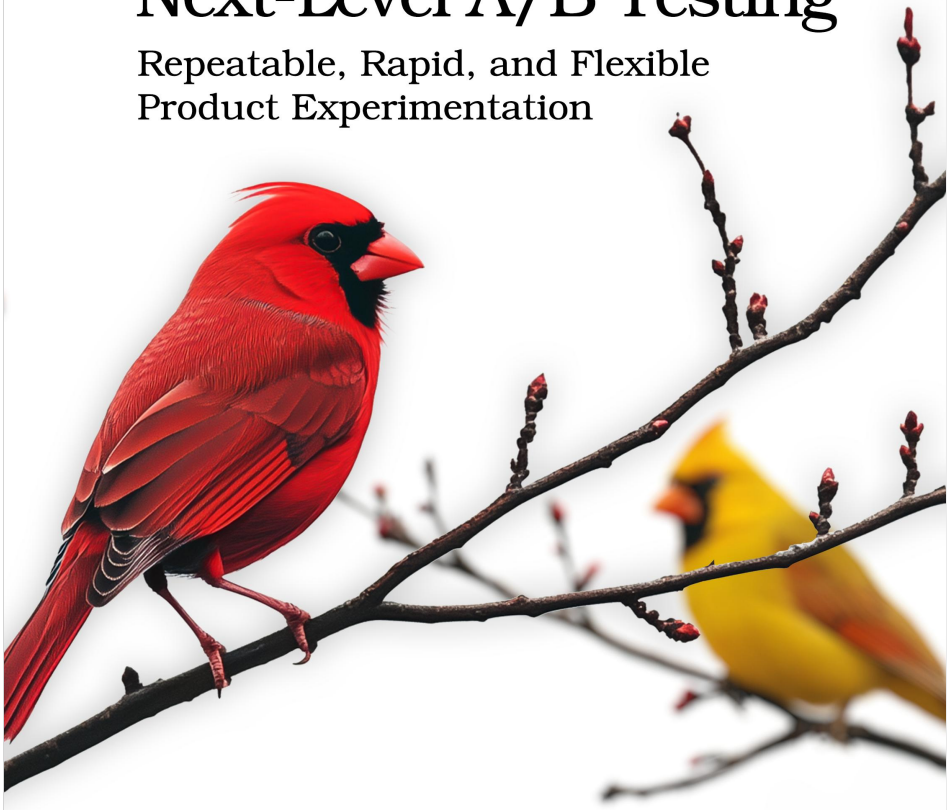
The Pragmatic Programmers

Next-Level A/B Testing

Repeatable, Rapid, and Flexible Product Experimentation



Leemay Nassery
edited by Vanya Wryter

This extract shows the online version of this title, and may contain features (such as hyperlinks and colors) that are not available in the print version.

For more information, or to purchase a paperback or ebook copy, please visit https://www.pragprog.com.

Tracking Metrics to Measure Experimentation Strategy Effectiveness

Imagine what could go wrong when running experiments on a product. You could misconfigure an experiment so users don't see the new feature in the test variant as expected. Or you could have a misconfigured control experience, which would invalidate the comparison you need to understand the effect of the test variant on the unchanged product experience. Worst-case scenario, you put ample time into configuring the test correctly before it's launched, but while the test runs, issues arise that lend themselves to inconclusive results caused by unbalanced variants or errors in your experimentation platform sampling logic. In all these examples, tools can be built to catch issues and improve the experiments launched on the platform.

The goal of any experimentation platform is to enable trustworthy A/B tests. A trustworthy experiment enables product decisions based on data instead of intuition or feelings. The more reliable insights are, the more knowledge your teams will have to improve user, business, and product metrics.

Before we dive into the tooling to verify experiments before launch and monitor active experiments, it helps to have it helps to have metrics that you aim to improve support the tooling implementation efforts. You want to show that you're maintaining and improving quality over time as teams evaluate more and more features with an A/B test. Key metrics aligning with this goal include: the count of aborted experiments, experiments meeting the gold standard, experiments yielding conclusive versus inconclusive results, and experiments that required rerunning after initially completing their duration. Let's explore each of these metrics in detail, starting with aborted experiments.

Usually, experiments are aborted early because of a misconfiguration. Sometimes, experiments can end early because metrics are so bad, but data scientists typically want to see the evaluation through the test duration to avoid novelty effects or seasonality concerns. The higher the number of conclusive results derived from tests, the more likely you've configured well-designed, trustworthy experiments.

If you notice that the number of experiments meeting the gold standard is similar to the number of aborted or inconclusive tests, consider redefining the requirements for meeting the gold standard on your experiment platform. This correlation should align with other metrics to some extent. If it doesn't, identifying the gaps between the gold standard and the issues leading to aborted tests or inconclusive results would be beneficial.

Metrics that may seem like good candidates to track but actually shouldn't be used to maintain experimentation quality are the number of product launches, number of features evaluated in the scope of an A/B test and then rolled out to all users, and number of experiments executed on the platform altogether. These metrics aren't ideal for tracking because they could be considered in most cases a vanity metric.

A vanity metric is a data point that sounds promising, but the value proposition is unclear and can't necessarily be tied to your platform. For instance, let's say you use the number of product launches to measure the quality of your experimentation platform. If product launches in the past quarter were low, that doesn't mean you're not running high-quality experiments on your platform or that the utilization of your platform has declined.

When you're brainstorming metrics to measure the effectiveness of your experimentation practices, consider the following questions to avoid using vanity metrics:

- Can the experimentation platform team take action to influence the metric?
- Does the metric tie into the experimentation platform's strategy and vision to continue to advance experimentation practices on the product?
- Is the value proposition from the perspective of teams using the experimentation platform to run A/B tests on the product reflected by the metric defintion?

If you can answer yes to the above questions, then you have a good a metric capable of tracking the quality of the experimentation platform over time.

To ensure quality of an A/B tests configuration is met, it helps to verify before launching. Let's see what's required to verify experiments before changes to the product are exposed to a subset of users.