Principles for Designing Reliable A/B Tests

Olle Marmenlind
ollemarm@kth.se

Course Code: LS1562

June 11, 2025

Abstract

This report reviews the fundamentals of designing and conducting reliable A/B tests for evaluating changes to products, services, or strategies. A/B testing is a common method for making data-driven decisions, but its usefulness depends on the reliability of the results. The report shows how unreliable results can lead to poor business decisions and wasted resources. It covers important topics such as the theoretical basis of A/B testing, including hypothesis testing and statistical analysis. It also examines practical aspects like choosing the right metrics (including an Overall Evaluation Criterion, OEC), determining the appropriate sample size and test duration to achieve statistical power, and performing correct random assignment and segmentation. The text also discusses common statistical problems and pitfalls, such as issues with multiple comparisons, misinterpretations of p-values and confidence intervals, and behavioral effects of novelty or user learning. The goal of this report is to provide a clear overview so that practitioners and students can create A/B tests that provide a stable and reliable basis for decision-making and optimization. A key conclusion is that reliability in A/B testing is complex, requiring a blend of statistical rigor, a well-thought-out methodology, an awareness of error sources, and an organizational culture that encourages sound experimentation.

A/B testing, experimental design, data analysis, hypothesis testing, statistical significance, product management, marketing optimization, variant testing, split testing.

Contents

1 Introduction

A/B testing, also known as split testing, is a method where two or more versions of an element are compared to determine which one performs best against a specific goal [1], [2]. It is a powerful tool for measuring the causal impact of changes, whether to a marketing email's subject line, a product's pricing model, or a backend recommendation algorithm. The method allows companies to make decisions based on empirical data by measuring how real users or systems actually behave in response to different variations. This can lead to improved performance, a better customer experience, and higher conversions [1].

Today, it is easier than ever for companies to conduct experiments due to the wide availability of A/B testing platforms and tools [1], [3]. Large technology companies like Microsoft, Google, and Amazon run thousands of A/B tests each year to continuously improve their products and services [1], [4], demonstrating the method's power. However, simply running an A/B test does not guarantee valid or useful results. The process is complex, with many potential pitfalls that can render the outcomes unreliable [4].

Reliability (or trustworthiness) in this context means that one can trust the experiment's results to be accurate and replicable under similar conditions. It also implies that the conclusions drawn are well-founded [5], [6]. An unreliable A/B test can lead to flawed business decisions. A change that is mistakenly identified as an improvement might actually harm performance. Conversely, a beneficial change could be rejected due to a poorly executed test. Both scenarios lead to wasted resources and can damage a company's revenue or reputation. The core problem is the gap between the ease of starting a test and the difficulty of ensuring its reliability. For teams lacking deep statistical expertise, these tools can create a false sense of being data-driven while decisions are actually based on flawed data.

The purpose of this report is, through an analysis of existing literature, to explain the key principles for designing reliable A/B tests. The report aims to answer the following questions:

1. What theoretical and statistical foundations are essential for understanding and achieving reliable A/B tests?
2. What practical steps and methods are critical when designing and conducting reliable A/B tests?
3. What are the most common pitfalls that threaten reliability, and how can they be avoided?

The report focuses on the general principles of A/B testing. While examples may be drawn from specific domains like web applications or marketing, the principles themselves are broadly applicable.

2 Theoretical Background/Fundamental Principles

To design and interpret A/B tests correctly, one must understand the underlying theoretical and statistical principles. This chapter explains A/B testing as a method, highlights the importance of metrics, and introduces the statistical concepts needed to assess the reliability of a result.

2.1 A/B Testing as a Controlled Experiment

A/B testing is fundamentally a controlled experiment designed to establish a causal relationship between a change (the independent variable) and an outcome (the dependent variable). In a typical A/B test, the current version, known as the control (version A), is compared with one or more modified versions, known as variants (version B, C, etc.) [2], [3].

A core principle for a reliable experiment is **randomization** [1], [2]. Subjects (e.g., users, visitors, recipients) are randomly assigned to groups that are exposed to either the control or a variant. This random assignment helps ensure that other confounding factors are distributed evenly across the groups. As a result, any observed difference in outcome can be attributed to the tested change with greater confidence [1], [7]. The unit being randomized—often a user, session, or device—is called the randomization unit [5]. The independent variable is the specific change being tested, while the dependent variable is the outcome measured to assess the effect.

2.2 Hypothesis Formulation

Every sound A/B test begins with a clear and testable hypothesis [1], [8]. A hypothesis is a proposed statement about the expected effect of a change on user behavior or system performance. A well-formed hypothesis is typically grounded in prior knowledge, such as customer feedback, data analysis, or established principles in a given field [1], [9].

Two opposing hypotheses are formulated:

- **Null Hypothesis (H0):** This hypothesis states that there is *no real difference* in the outcome between the control and the variant. Any observed difference is attributable to random chance [1], [8].
- **Alternative Hypothesis (H1 or HA):** This hypothesis states that there is a *real difference* in the outcome, caused by the tested change [1], [8]. It can be directional (e.g., "variant B will achieve a higher open rate") or non-directional ("there will be a difference in open rates between A and B").

For example, a marketing hypothesis could be: "If we use an emoji in the email subject line, the open rate will increase by at least 5%, because it will make the email stand out in a crowded inbox."

2.3 Key Metrics

The selection of metrics is critical to a test's relevance and reliability. Metrics must be directly tied to the hypothesis and broader business objectives. The choice of metrics is domain-specific.

- In **marketing**, metrics might include email open rates, click-through rates (CTR), or lead generation conversions.
- In **product management**, metrics could be user engagement, feature adoption rates, user retention, or task success rates.
- For **backend changes**, relevant metrics might be server response time, processing efficiency, or error rates.

To prevent optimizing one metric at the expense of others, many experts recommend using an **Overall Evaluation Criterion (OEC)** [1], [2]. An OEC is a single, quantifiable score—often a composite of several metrics—that reflects long-term strategic goals (e.g., customer lifetime value, net profit per user). Defining a good OEC is challenging but essential. It forces an organization to clarify what "success" truly means and prevents A/B tests from chasing short-term gains that are detrimental overall. For instance, an e-commerce company might define its OEC as a weighted sum of average order value and purchase frequency, minus a cost associated with customer churn.

2.4 Fundamental Statistical Concepts for Reliability

Statistics form the core of A/B testing. The following concepts are crucial for determining whether an observed result is trustworthy or merely a product of random chance.

- **Significance Level (α):** Before the test, a significance level, α (alpha), is defined. This is the acceptable probability of making a Type I error—concluding there is a difference when one does not actually exist. A common value for α is 0.05 (5%), meaning a 5% risk of a false positive [1], [8], [9].
- **P-value:** After data collection, a p-value is calculated. The p-value is the probability of observing a result at least as extreme as the one measured, assuming the null hypothesis is true [8], [9]. If the p-value is less than α (e.g., $p < 0.05$), the result is deemed **statistically significant**, and the null hypothesis is rejected. It's crucial not to misinterpret the p-value as the probability that the null hypothesis is true. A low p-value simply indicates that the observed data is inconsistent with the null hypothesis.
- **Statistical Power (1-β):** Power is the probability of detecting a real effect if it exists, thereby avoiding a Type II error (failing to detect a real difference). Power is denoted as 1-β, where β is the probability of a Type II error. A power of 80% or higher is a common target. Low power, often due to an insufficient sample size, increases the risk of incorrectly concluding "no difference exists."

- **Confidence Interval (CI):** A confidence interval provides a range of plausible values for the true effect size (e.g., the true difference in conversion rates). A 95% confidence interval suggests that if the experiment were repeated many times, 95% of the calculated intervals would contain the true effect size. A wide interval indicates high uncertainty, whereas a narrow one suggests greater precision. If the interval includes zero, the result is typically not statistically significant.
- **Effect Size:** This metric quantifies the magnitude of the difference between the groups. While statistical significance indicates if an effect is likely real, effect size tells you if it is large enough to be practically meaningful. A tiny, unimportant effect can become statistically significant with a massive sample size, so it is vital to consider both significance and effect size.

A fundamental tension exists between the agile demand for rapid results and the time required for statistically sound A/B tests. Experiments must run long enough with a sufficient sample size to achieve adequate power [1], [8], [10]. This necessitates careful planning to ensure statistical requirements are understood and met before launching a test.

2.5 Validity and Reliability in A/B tests

For the results of an A/B test to be trustworthy, the experiment must be both valid and reliable. **Validity** addresses whether the test accurately measures what it intends to and whether the conclusions are generalizable [6], [10], [11], [12]. Key types of validity include:

- **Internal validity:** Confidence that the tested change caused the observed effect, not a confounding variable. Proper randomization is essential.
- **External validity (generalizability):** The extent to which results can be applied to other populations, settings, or times.
- **Construct validity:** Whether the chosen metrics accurately represent the abstract concepts being studied (e.g., "customer satisfaction" or "engagement").
- **Statistical conclusion validity:** The correctness of the statistical inferences. Threats include low power and violations of statistical assumptions.

Reliability refers to the consistency and repeatability of the measurements. If the experiment were repeated under identical conditions, would it yield the same results? Ensuring both validity and reliability is central to conducting trustworthy A/B tests.

3 Methodological Aspects of A/B-test Design for Reliability

This chapter covers the practical and methodological steps for designing and executing reliable A/B tests. It emphasizes careful goal setting, metric selection, sample size determination, proper randomization, and awareness of common pitfalls. Despite the accessibility of A/B testing tools [1], deep methodological and statistical understanding is essential to avoid producing unreliable results and a false sense of being data-driven [4], [5], [6].

3.1 Goal Definition and Metric Selection

A reliable A/B test begins with clear goals tied to business objectives [2], [8]. Without clear goals, it is impossible to form a strong hypothesis and select appropriate metrics.

- **Primary metric:** The single metric that directly measures the success of the hypothesis. This should be the most important outcome you aim to influence.
- **Secondary metrics:** These provide additional context, helping to explain *why* a change produced a certain effect. They can also capture unintended positive or negative side effects.
- **Guardrail metrics:** Critical metrics monitored to ensure that improvements in the primary metric do not negatively impact other vital areas [11]. For example, a change might increase conversions (primary) but also increase customer support tickets (guardrail).
- **Overall Evaluation Criterion (OEC):** As previously mentioned, an OEC is a strategic tool for aligning A/B tests with long-term goals [1], [5], [11]. It helps prevent local optimization that could be detrimental to the business as a whole.

3.2 Sample Size and Statistical Power

To be reliable, an A/B test must have sufficient statistical power to detect a meaningful effect if one exists [1], [8], [10]. The sample size (the number of subjects in each group) is the primary driver of power. The required sample size should be calculated *before* the test begins, based on:

1. **Baseline conversion rate:** The current performance of the primary metric.
2. **Minimum Detectable Effect (MDE):** The smallest effect size that is considered practically significant and that you want the test to be able to detect.
3. **Significance level (α):** Typically 0.05.
4. **Desired Statistical Power (1-β):** Typically 80% or higher.

Numerous online calculators are available for this purpose. Running a test with an inadequate sample size leads to low power, increasing the risk of a Type II error (a false negative) [10].

3.3 Test Duration and Exposure

Test duration is linked to sample size and the volume of traffic or subjects available. The test must run long enough to achieve the pre-calculated sample size [1], [2].

Key considerations include:

- **Cyclical patterns:** Behavior can vary significantly by day of the week or time of day. To account for this, tests should ideally run for full weekly cycles—often one or two full weeks [1], [8].
- **Seasonality:** Longer test durations may be necessary to average out the effects of holidays, marketing campaigns, or other external events [1], [8].
- **Novelty and Learning Effects:** Users may initially react positively to a change simply because it's new (novelty effect) or perform worse as they adapt to it (learning effect). Both phenomena can mask the true long-term impact, sometimes requiring longer experiments to get a reliable reading [1], [8].

3.4 Randomization and Allocation

Correct randomization is the cornerstone of a reliable A/B test. Subjects must be randomly assigned to the control and variant groups [1], [2]. Important considerations:

- **Randomization unit:** Typically users, identified via cookies, user IDs, or customer accounts [1].
- **Consistent exposure:** An assigned subject should consistently see the same version throughout the test to avoid data contamination [1], [8].
- **Allocation ratio:** A 50/50 split between control and one variant generally provides maximum statistical power [10].
- **Sample Ratio Mismatch (SRM):** A significant deviation from the intended allocation (e.g., getting a 40/60 split instead of 50/50) is a major red flag. SRM often indicates a bug in the randomization or data logging process and can invalidate the entire test [1].

3.5 Segmentation

Analyzing results across different segments (e.g., new vs. returning customers, users on different subscription plans, or geographic regions) can yield deeper insights than looking only at the overall average [2], [8], [11]. However, excessive post-hoc segmentation carries a high risk. Searching for significant results across many small groups (a practice known as *p-hacking*) inflates the probability of finding a false positive (Type I error) due to the multiple comparisons problem [4], [11]. Segments for analysis should ideally be defined in advance.

3.6 Handling Multiple Comparisons and "Peeking"

When testing multiple variants against a control or tracking numerous metrics, the problem of **multiple comparisons** arises [4], [11]. Each comparison has its own risk of a Type I error. As the number of comparisons grows, the overall probability of at least one false positive increases. Statistical adjustments like the Bonferroni correction can be used to control for this, but they also reduce statistical power [2], [8]. Another common threat is **"peeking"** at the results and stopping the test as soon as statistical significance is reached [1], [4]. This practice dramatically increases the Type I error rate, as random fluctuations can easily produce a transient significant result. The test duration should be determined beforehand based on the sample size calculation.

3.7 Common Pitfalls and Best Practices in A/B tests

Below are some common pitfalls and best practices for general A/B testing.

3.7.1 Common Pitfalls:

- **Testing too many changes at once:** In a multivariate test, it's impossible to isolate which specific change caused the observed effect [8], [9].
- **Ignoring practical significance:** A result can be statistically significant but too small to be meaningful for the business.
- **Implementation flaws:** Technical issues, such as bugs in one variant, slow page load times, or incorrect email delivery, can corrupt results [1], [4]. Thorough quality assurance is essential [8].
- **External events:** Competitor actions, news cycles, or technical outages can influence user behavior and confound results [2], [8], [9].
- **Poor documentation:** Failure to document hypotheses, designs, and outcomes makes it impossible to build institutional knowledge and learn from past mistakes [9], [11].

3.7.2 Best Practices:

- **Isolate variables:** Test one change at a time to establish clear causality [2], [8], [9].
- **Formulate strong, data-driven hypotheses:** Don't test random ideas. Ground hypotheses in prior research and data [1], [2].
- **Run tests to completion:** Adhere to the pre-calculated sample size and test duration [2], [8], [9].
- **Validate the testing system and results:**

  - **A/A tests:** Run tests where both versions are identical to verify that the system is working correctly. You should see a statistically significant result only about 5
  - **Replicate important wins:** If a result is surprising or has major business implications, consider running a replication study to confirm the findings.

- **Build a knowledge repository:** Document and share all experiment results—both successes and failures—to foster organizational learning [5], [11].

Managing interactions between concurrent experiments is a significant challenge [4], [10], [11]. If multiple tests affect the same users and metrics, their effects can become entangled, making it difficult to attribute outcomes correctly. Solutions include designing orthogonal experiments or creating separate user layers for different tests.

4 Discussion

This report has outlined the core principles for designing and executing reliable A/B tests. A clear takeaway is that while A/B testing is a powerful tool, its value is entirely dependent on the rigor applied at every stage of the process. The key pillars of reliability are a solid understanding of hypothesis testing, thoughtful selection of metrics (especially a strategic OEC), rigorous sample size calculation to ensure statistical power, and proper randomization. Equally important is an awareness of statistical pitfalls, including multiple comparisons, p-value misinterpretation, and behavioral biases like novelty effects.

These components are deeply interconnected. A brilliant hypothesis [2], [8] is useless if the test lacks statistical power due to an insufficient sample size [10]. Perfect randomization [1] can produce precise results that are nonetheless misleading if the wrong metrics were chosen or if practical significance is ignored [8]. Reliability, therefore, emerges at the intersection of sound theory, statistics, and methodology.

A major challenge is measuring the long-term effects of a change using A/B tests, which are often short-term in nature [1], [10]. A test might show a short-term lift in a metric, but it is difficult to know how the change impacts long-term customer loyalty or brand perception. Identifying and validating proxy metrics that are predictive of long-term outcomes remains an active area of research.

Ethical considerations are also paramount. There is a risk that A/B testing can be used to optimize for manipulative practices—such as deceptive marketing copy or confusing pricing structures—that exploit user psychology [13]. Such tactics may produce positive results in a short-term A/B test but can erode user trust and damage the brand in the long run. Therefore, reliable A/B testing must incorporate an ethical dimension, aiming to create genuine value for the user, not just to optimize numbers.

Perhaps the most critical—and most difficult—component is fostering an organizational culture that supports reliable experimentation [5], [6], [10], [11]. Even with the best tools, A/B testing can become unreliable if the organization does not value honest results (including null and negative outcomes) and is unwilling to invest the necessary time and resources. A culture that pressures teams to produce quick, positive results can encourage poor practices like peeking, p-hacking, or ignoring red flags, thereby undermining the entire process.

Future trends in A/B testing point toward greater use of artificial intelligence (AI) and machine learning for tasks like automated hypothesis generation, variant design, and even user simulations to accelerate testing [7], [10]. This technology creates new opportunities but also introduces new reliability challenges, such as the need to validate and interpret AI-generated insights [7].

Ultimately, there is often a conflict between optimizing for short-term, easily measured metrics and achieving long-term goals related to customer satisfaction and brand loyalty [1], [10]. A/B tests excel at measuring immediate behavioral responses. However, if they are used solely to chase short-term gains, they can inadvertently degrade the long-term customer experience. A holistic approach that balances quantitative A/B test data with qualitative insights is often necessary. In conclusion, achieving reliable A/B tests requires a comprehensive approach that integrates statistical expertise, methodological discipline, and a supportive organizational culture.

5 Conclusion

This report has detailed the essential principles for conducting reliable A/B tests. The analysis shows that a trustworthy test is not an accident but the result of deliberate choices and a rigorous process. The report has addressed the three initial research questions, with the conclusions summarized below:

The first question concerned the essential theoretical and statistical foundations for reliability. The answer is that a practitioner must have:

- **A strong theoretical and statistical foundation:** A firm grasp of experimental design, hypothesis testing, statistical power, p-values, confidence intervals, and effect sizes is non-negotiable. Misunderstanding these concepts frequently leads to flawed conclusions.

The second question addressed the critical practical steps and methods in test design and execution. The answer can be summarized by two key rules:

- **A clear metric strategy:** Defining clear goals and relevant metrics, including an Overall Evaluation Criterion (OEC) tied to long-term objectives, is crucial for drawing meaningful and trustworthy conclusions.
- **Rigorous test design:** Careful calculation of the required sample size, selection of an appropriate test duration, and ensuring proper, unbiased randomization are

essential to minimize the risk of erroneous results.

The third and final question was about the most common pitfalls and how to avoid them. To guard against invalid results, one must focus on two areas:

- **Awareness of pitfalls and adherence to best practices:** Actively avoiding common mistakes like testing too many changes at once, misinterpreting p-values, or "peeking" at results is critical. Adherence to established best practices is paramount.
- **Organizational culture and ethics:** Ultimately, the organization's commitment is the most important factor. A culture that values rigor, embraces honest results (including failures), and upholds ethical principles is what enables A/B testing to deliver true, sustainable value.

The central thesis is that while the mechanics of A/B testing may seem simple, conducting them reliably requires significant expertise and discipline. There is a profound difference between "running an A/B test" and "running a *reliable* A/B test."

References

[1] R. Kohavi, D. Tang, and Y. Xu, *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* Cambridge, UK: Cambridge University Press, 2020.

[2] J. Lazar, J. H. Feng, and H. Hochheiser, *Research Methods in Human-Computer Interaction*, 2nd. Cambridge, MA, USA: Morgan Kaufmann, 2017.

[3] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann, "Online controlled experiments at large scale," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*, Chicago, IL, USA, Aug. 2013, pp. 1168–1176. DOI: 10.1145/2487575.2488217.

[4] T. Crook, R. Kohavi, R. Longbotham, and B. Frasca, "Seven pitfalls to avoid when running controlled experiments on the web," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*, Paris, France, Jun. 2009, pp. 1105–1114. DOI: 10.1145/1557019.1557139.

[5] X. Amatriain and J. Basilico. "Netflix recommendations: Beyond the 5 stars (part 1)." [Online], Accessed: Jun. 11, 2025. [Online]. Available: https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429.

[6] A. M. Cirucci and U. M. Pruchniewska, *UX Research Methods for Media and Communication Studies: An Introduction to Contemporary Qualitative Methods.* New York, NY, USA: Routledge, 2023.

[7] J. W. G. Addo, A. M. E. Gyamfi, E. A. Adu-Gyamfi, E. T. Tchao, and P. K. O. Asante. "Agenta/b: A llm agent-based a/b testing system for real web applications." arXiv preprint. arXiv: 2504.09723. [Online]. Available: https://arxiv.org/abs/2504.09723.

[8] F. Quin, D. Weyns, M. Galster, and C. M. da Costa Silva, "A/b testing: A systematic literature review," *Journal of Systems and Software*, vol. 197, p. 111 093, Mar. 2023. DOI: 10.48550/arXiv.2308.04929.

[9] L. Wasserman, A. Ramdas, and S. Balakrishnan, "Universal inference," *Proceedings of the National Academy of Sciences*, vol. 117, no. 29, pp. 16 880–16 890, Jul. 2020. DOI: 10.1073/pnas.1922664117.

[10] R. Kohavi, D. Tang, Y. Xu, L. G. Hemkens, and J. P. A. Ioannidis, "Online randomized controlled experiments at scale: Lessons and extensions to medicine," *Trials*, vol. 21, no. 1, p. 150, Feb. 2020. DOI: 10.1186/s13063-020-4084-y.

[11] A. Fabijan, P. Dmitriev, B. Arai, A. Drake, S. Kohlmeier, and A. Kwong, "A/b integrations: 7 lessons learned from enabling a/b testing as a product feature," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, Melbourne, Australia, May 2023, pp. 773–785. DOI: 10.1109/ICSE-SEIP58684.2023.00033.

[12] A. M. Lund, "Measuring usability with the use questionnaire," *Usability Interface*, vol. 8, no. 2, pp. 3–6, Oct. 2001.

[13] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, “The dark (patterns) side of ux design,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’18, New York, NY, USA: Association for Computing Machinery, 2018, 534:1–534:14, ISBN: 9781450356206. DOI: 10.1145/3173574.3174108. [Online]. Available: https://doi.org/10.1145/3173574.3174108.