Principles for Designing Reliable A/B Tests for Evaluating User Interface Changes in Web Applications

Olle Marmenlind
ollemarm@kth.se
Course Code: LS1562

June 11, 2025

Abstract

This report reviews the fundamentals of how to design and conduct reliable A/B tests when evaluating changes in user interfaces (UI) for web applications. A/B testing is a common method today for making data-driven decisions in web development, but the method's usefulness depends on the reliability of the results. The report shows how unreliable results can lead to poor design decisions and a waste of resources. The report covers important topics such as the theoretical basis for A/B testing, for example, hypothesis testing and statistical analysis. The report also looks at practical aspects like choosing the right metrics (including an overall evaluation criterion, OEC), determining the right sample size and test duration to achieve statistical power, and how to perform a correct random assignment and segmentation. The text also discusses common statistical problems and pitfalls. These can include issues with multiple comparisons, misinterpretations of p-values and confidence intervals, and effects of novelty or user learning. The goal of the report is to provide a clear overview so that students and developers can create A/B tests that provide a stable and reliable basis for decisions about UI improvements. An important conclusion is that reliability in A/B tests is complex. It requires a mix of statistical rigor, a well-thought-out methodology, an awareness of sources of error, and a company culture that encourages good experimentation.

A/B testing, user interface, web applications, reliability, nudging, applied behavioral science, product design, data analysis, variant testing, split testing.

Contents

1 Introduction

A/B testing, also known as split testing, is a method where two or more versions of a webpage or feature are compared to see which one performs best towards a specific goal [1], [2]. For web applications, A/B testing is a good tool for measuring the effect of changes in the user interface (UI). This could involve a new layout, different colors, or new text on a button [1]. The method allows companies to make decisions based on data because it measures how real users actually behave with different designs. This can, in turn, lead to a better user experience (UX) and more conversions [1].

Today, it is easier than ever for companies to conduct experiments because there are many available tools for A/B testing [1], [3]. Large tech companies like Microsoft, Google, and Amazon run thousands of A/B tests each year to continuously improve their products and services [1], [4]. This shows how powerful the method can be. However, simply running an A/B test does not mean you will get good or correct answers. It is actually quite complicated and there are many pitfalls that can make the results unreliable [4].

Reliability (or trustworthiness) in this context means that one can trust that the experiment's results are correct and can be replicated under similar conditions. It also means that the conclusions drawn are well-founded [5], [6]. An unreliable A/B test can lead to incorrect design decisions. A UI change that is mistakenly seen as an improvement may actually worsen the user experience. At the same time, a truly good change could be rejected due to a poor test. Both cases lead to a waste of development resources and can damage the company's reputation or revenue. The problem is thus the difference between how easy it is to start a test and how difficult it is to ensure that it is reliable. Especially for teams without deep statistical knowledge, all the tools can create a false sense of being data-driven, when in fact the decisions are based on bad data.

The purpose of this report is to, by analyzing existing literature, explain the most important principles for designing reliable A/B tests for UI changes in web applications. The report aims to answer the following questions:

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

The report focuses on principles for A/B testing of UI changes in web applications. Many principles are general, but the emphasis is not on mobile apps or other areas, unless they illustrate a general point.

2 Theoretical Background/Fundamental Principles

To be able to design and interpret A/B tests properly, one must understand the theoretical and statistical principles that underlie them. This chapter explains A/B testing as a method, highlights important metrics, and introduces the statistical concepts needed to assess whether a result is reliable.

2.1 A/B Testing as a Controlled Experiment

A/B testing is fundamentally a controlled experiment. The goal is to see if there is a causal relationship between a change (independent variable) and an outcome (dependent variable). When testing UI changes in web apps, one usually compares the current design, the control version (version A), with one or more modified versions, called variants (version B, C, and so on) [2], [3].

A fundamental principle for a reliable experiment is randomization [1], [2]. Users visiting the web app are randomly divided into groups that either see the control version or a variant. This random assignment ensures that other factors that could affect the outcome are spread evenly between the groups. Then, one can say with greater certainty that a difference in results is due to the tested UI change [1], [7]. The unit that is randomized, usually a user identified by a cookie, is called the randomization unit [5]. The independent variable is the specific UI change being tested, for example, a new button color. The dependent variable is the outcome measured to see the effect, for example, the click-through rate.

2.2 Hypothesis Formulation

Every good A/B test starts with a clear and testable hypothesis [1], [8]. A hypothesis is an assumption about how a UI change is expected to affect users and metrics. A good hypothesis is often based on prior knowledge from user studies, analytics data, or known UX principles [1], [9].

Two opposing hypotheses are usually set up:

- **Null Hypothesis (H0):** This hypothesis states that there is *no real difference* in outcome between the control version and the variant. Any observed differences are due to chance alone [1], [8].
- **Alternative Hypothesis (H1 or HA):** This hypothesis states that there is a *real difference* in outcome. The difference is caused by the tested UI change [1], [8]. The alternative hypothesis can either be directional (e.g., "variant B will have a higher conversion rate") or non-directional (e.g., "there is a difference in conversion rate between A and B").

An example of a hypothesis could be: "If we change the text on the buy button from 'Add to Cart' to 'Buy Now', the click-through rate will increase by at least 10

2.3 Key Metrics

The choice of metrics is very important for the test's relevance and reliability. The metrics must be directly linked to the hypothesis and the overall goals. Common metrics when evaluating UI changes in web applications are [1], [8]:

- **Conversion Rate:** The percentage of users who perform a desired action, such as making a purchase.
- **Click-Through Rate (CTR):** The percentage of users who click on a specific element, such as a link or a button [1], [8].
- **Bounce Rate:** The percentage of users who leave the website after viewing only a single page.
- **Average Session Duration:** The average time a user is active on the website.
- **Task Success Rate:** The percentage of users who successfully complete a specific task.

To avoid a situation where an improvement in one metric leads to undesirable negative effects on another, many experts recommend using an Overall Evaluation Criterion (OEC) [1], [2]. An OEC is a single, measurable value, often a combination of several metrics, that reflects the company's long-term goals (such as customer lifetime value or total revenue per user). Defining a good OEC is difficult but important. It forces the organization to decide what "success" really means. Without a good OEC, there is a risk that A/B tests will only optimize for short-term gains that do not benefit the whole, causing a loss of confidence in the experiments. An e-commerce company, for example, could define its OEC as a weighted sum of average order value and number of completed purchases per user, minus a cost for canceled subscriptions. This would balance short-term sales against long-term customer loyalty.

2.4 Fundamental Statistical Concepts for Reliability

Statistics are at the core of A/B testing. The following concepts are important for determining whether a result is reliable or just due to chance.

- **Significance Level (α):** Before the test begins, a significance level, called α (alpha), is determined. This is the risk you accept of concluding that there is a difference when there actually isn't one (a Type I error). α is usually set to 0.05 (5
- **P-value:** After the test is complete and data has been collected, a p-value is calculated. The p-value is the probability of obtaining a result at least as extreme as the one observed, assuming the null hypothesis is true [8], [9]. If the p-value is lower than α (e.g., $p < 0.05$), the result is said to be statistically significant, and the null hypothesis is rejected. It is important to interpret the p-value correctly. Many people mistakenly believe that the p-value is the probability that the null hypothesis is true, or the probability that the result is due to chance. A low p-value

only means that the data is unlikely if the null hypothesis is true. It says nothing about the size or importance of the effect. Just staring at the p-value is a common pitfall.

- **Statistical Power (1-β):** Statistical power is the probability of detecting a real effect if one exists [1], [8]. Power is often written as 1-β (beta), where β is the risk of a Type II error (missing a real effect). The aim is often for a power of at least 80
- **Confidence Interval (CI):** A confidence interval is a range of values where the true effect (e.g., the real difference in conversion) is likely to lie, with a certain level of confidence (often 95
- **Effect Size:** The effect size is a measure of how large the difference is between the groups [1], [8]. Statistical significance (the p-value) only tells you if the difference is likely real. The effect size tells you how large and practically relevant it is. A small, insignificant effect can become statistically significant if you have a very large number of participants. Therefore, you must look at both statistical significance and effect size to make good decisions.

There is a conflict between the need for quick results in agile development and the time required to conduct statistically reliable A/B tests. The tests must run long enough with enough participants to achieve statistical power [1], [8], [10]. This leads to development teams either running tests that are too short, which lowers reliability, or testing becomes a bottleneck in development. A fundamental principle is therefore to plan and understand the statistical requirements before starting an A/B test.

2.5 Validity and Reliability in A/B tests

For the results of an A/B test to be reliable, it must be both valid and reliable. **Validity** is about whether the test measures what it intends to measure and whether the conclusions are correct and can be generalized [6], [10], [11], [12]. There are different types of validity:

- **Internal validity:** How sure can you be that it was the UI change that caused the effect, and not something else? Good randomization is crucial for internal validity.
- **External validity (or generalizability):** How well can the results from the test apply to other groups, situations, or times? This is affected by whether the test participants are representative and whether the test environment resembles the real one.
- **Construct validity:** This concerns how well the metrics and the UI change actually represent the theoretical concepts one wants to investigate (like "usability" or "engagement").
- **Statistical conclusion validity:** This concerns whether the statistical conclusions are correct. Low statistical power or incorrect use of statistical methods are

threats to this validity.

Reliability of the measurements is about how consistent and repeatable the results are. If you were to repeat the test under the exact same conditions, would you get the same result? This is affected by how accurate the measurement tools are. Ensuring the different types of validity is central to reliability. To strengthen external validity, for example, one must consider whether the test group is representative of all users. Internal validity is threatened if the groups differ in other ways than the UI change being tested.

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

This chapter focuses on the practical and methodological steps that are crucial for designing and conducting reliable A/B tests. It involves careful selection of goals and metrics, determination of sample size and test duration, correct randomization, and how to handle segmentation and common problems. Although A/B testing tools are easy to access [1], it is important to have deep methodological and statistical knowledge. Otherwise, you risk getting results that are not reliable and that only provide a false sense of being data-driven [4], [5], [6].

3.1 Goal Definition and Metric Selection

A reliable A/B test starts with clear goals that are linked to business or usability objectives [2], [8]. Without clear goals, it is difficult to create a good hypothesis and choose the right metrics.

- **Primary metrics:** The single metric that directly measures whether the test's hypothesis was successful. It should be the most important outcome you want to influence.

- **Secondary metrics:** These metrics can provide additional insights and help to understand why a change had a certain effect. They can also capture other positive or negative effects.

- **Guardrail metrics:** Critical metrics that are monitored to ensure that an improvement in the primary metric does not come at the expense of something else important, like the user experience [11]. For example, a change might increase the click-through rate (primary) but also increase loading time (guardrail).

- **Overall Evaluation Criterion (OEC):** As mentioned earlier, the OEC is a strategic tool to ensure that A/B tests work towards long-term goals [1], [5], [11]. It helps to avoid optimizing for short-term gains in a single metric that could harm the overall picture. Defining a good OEC is difficult but very important for A/B tests to yield truly valuable and reliable results.

3.2 Sample Size and Statistical Power

For an A/B test to provide a reliable result, it must have sufficient statistical power to be able to find a meaningful difference if one exists [1], [8], [10]. The sample size, i.e., the number of users in each group, is absolutely crucial for the test's power. The required sample size should be calculated before the test begins, based on the following factors [1], [2], [8]:

1. **Baseline value:** The current value of the primary metric.
2. **Minimum Detectable Effect (MDE):** The smallest difference between control and variant that you consider practically important and want to be able to detect. A too small MDE requires a very large sample.
3. **Significance level (α):** Usually 0.05.
4. **Desired Statistical Power (1-β):** Usually 80

There are many online calculators for calculating sample size. Running a test with too small a sample results in low statistical power. The risk is then high that you miss a real effect (Type II error) and mistakenly believe that the UI change made no difference [10].

3.3 Test Duration and Exposure

The test duration is related to the sample size and how much traffic you have per day or week. The test must run long enough to reach the calculated number of observations [1], [2].

One should also consider:

- **Weekly cycles:** User behavior can differ greatly between weekdays and weekends. To get a fair picture, tests should be run in full weeks, for example, at least one or two [1], [8].
- **Seasonal variations:** Longer tests may be needed to even out effects from seasons, campaigns, or other external events that can affect behavior [1], [8].
- **Novelty effects:** Users may initially react positively to a new design just because it is new. The effect may decrease as they get used to it [1], [8].
- **Learning effects:** Users may initially perform worse with a new design because they need time to learn it, even if it is better in the long run [1], [8]. Both of these effects may require longer tests to get a reliable picture of the long-term effect.

3.4 Randomization and Allocation

Correct randomization is the foundation of a reliable A/B test. Users must be randomly assigned to control and experimental groups [1], [2]. Important things to consider:

- **Randomization unit:** Usually individual users, identified via cookies or user IDs [1].
- **Consistent exposure:** A user who has been assigned to a group should always see the same version throughout the test to avoid "contaminating" the data, meaning the user's behavior is influenced by exposure to multiple variants, which leads to invalid measurement results. [1], [8].
- **Allocation ratio:** Usually, the traffic is split 50/50 between the control and one variant to achieve maximum statistical power [10].
- **Sample Ratio Mismatch (SRM):** If the distribution of users becomes very uneven (e.g., 35/65 instead of 50/50), it is a strong warning sign. This is called SRM and indicates problems with randomization or data collection, which can invalidate the test [1].

3.5 Segmentation

Analyzing A/B test results for different user segments (such as new vs. returning users, or mobile vs. desktop) can provide deeper insights than just looking at the average [2], [8], [11]. A UI change can be good for one group but bad for another. But too much segmentation is a risk. Searching for significant results in small groups after the fact (called *p-hacking*) increases the risk of false positive results (Type I error). This is due to the problem of multiple comparisons [4], [11]. One should preferably decide which segments to analyze in advance.

3.6 Handling Multiple Comparisons and "Peeking"

When testing multiple variants against a control, or measuring many different things at the same time, the problem of multiple comparisons arises [4], [11]. Each individual test has a risk of a Type I error (e.g., 5Another common threat is "peeking," which is constantly checking the results and stopping the test as soon as you see a statistically significant difference [1], [4]. This increases the risk of false positive results, as random variations early in a test can look like real effects. Instead, one should determine the test duration in advance. Alternatively, one can use more advanced methods that allow for continuous monitoring without increasing the error rate [9].

3.7 Common Pitfalls and Best Practices in UI A/B tests

Here are some specific pitfalls and best practices for A/B tests of UI.

3.7.1 Common Pitfalls:

- **Testing too many changes at once:** If several things are changed in a variant, it is impossible to know which change caused the effect [8], [9].

- **Ignoring small effects:** Even if an effect is not statistically significant, small, consistent effects over several tests can be practically important.
- **Misinterpreting statistical significance:** Believing that statistical significance is the same as practical importance is a common error [8]. A small, irrelevant difference can become significant with enough data.
- **Influence of external events:** Major news, technical problems, or competitors' campaigns can affect test results [2], [8], [9].
- **Technical issues:** A variant may load slower or have bugs, which ruins the results [1], [4]. It is important to quality assure the test thoroughly [8].
- **Poor documentation:** If you do not document hypotheses, design, and results, it is difficult to learn and avoid making the same mistakes again [9], [11].

3.7.2 Best Practices:

- **Test one thing at a time:** To be able to link the effect to the correct change, one should isolate variables [2], [8], [9].
- **Have a clear, data-driven hypothesis:** Do not test random ideas. Hypotheses should be based on insights [1], [2].
- **Run tests long enough:** Respect the calculated sample sizes and test durations [2], [8], [9].
- **Validate results:**
  - **A/A tests:** Run tests where both versions are identical to check that the experiment system is working correctly. You would then expect significant results in only about 5
  - **Repeat winning tests:** If a result is surprising or has a large impact, a repetition can increase confidence.
- **Document and share:** Create a knowledge base of experiments, both successful and unsuccessful, to promote learning [5], [11].
- **Segment with caution:** Use segments to gain deeper insights, but be aware of the risk of false positives.

Managing interactions between experiments is a major challenge when multiple A/B tests are run simultaneously [4], [10], [11]. If two tests affect the same user and metric, their effects can be mixed, making it difficult to see the effect of a single test. Solutions can be to design tests so they are independent of each other or to limit which tests are run at the same time on the same user.

4 Discussion

The report has reviewed a range of principles that are important for designing and conducting reliable A/B tests of UI changes. A clear insight is that although A/B testing is a powerful tool, its value depends entirely on how carefully one is at every step of the process. The most important principles are a good understanding of hypothesis testing, the right choice of metrics (especially the strategic OEC), careful calculation of sample size to achieve statistical power, and correct randomization. It is also fundamental to be aware of statistical pitfalls such as problems with multiple comparisons, misinterpretation of p-values, and effects like the novelty effect.

All these parts are interconnected. A good hypothesis [2], [8] is worthless if the test lacks statistical power due to too few participants [10]. A correct randomization [1] can yield accurate results that are still misleading if the wrong metrics have been chosen or if one ignores the practical significance [8]. Reliability is thus created where theory, statistics, and methodology meet.

A major challenge is to measure the long-term effects of UI changes with A/B tests that are often short [1], [10]. A test can show a quick increase in clicks, but it is difficult to know how the change affects customer loyalty in the long run. Finding and validating metrics that can predict long-term results is an active area of research.

Another challenge is to manage interactions between different experiments running simultaneously [4], [10], [11]. In a dynamic web environment where multiple teams are testing in parallel, it is difficult to isolate the effect of a single change. Ethical issues in experiments are also important. There is a risk that A/B testing is used to optimize manipulative designs, so-called "dark patterns," that trick users into doing things they do not want to do [13]. Such patterns may look good in an A/B test in the short term, but can damage users' trust and the brand in the long term. Reliable A/B testing should therefore also have an ethical dimension, where the goal is not just to optimize numbers, but to do so with respect for the user.

Building a culture that supports reliable experiments is perhaps the most important, but also the most difficult, part [5], [6], [10], [11]. Even with the best tools, A/B tests can become unreliable if the organization does not value honest results (even negative ones) and is not willing to invest time and resources. A culture that pushes for quick, positive results can lead to peeking at data, "p-hacking," or ignoring warning signs, which destroys reliability.

Future trends in A/B testing point towards more use of artificial intelligence (AI) and machine learning. This could involve automatic generation of hypotheses, design of variants, and even simulation of users to speed up tests [7], [10]. This technology offers new possibilities, but also new challenges for reliability, such as the need to validate AI-generated insights [7].

There is a conflict between optimizing for short-term, easily measured goals and the long-term goals of user experience (UX) and loyalty [1], [10]. A/B tests are good at measuring immediate behaviors. But UX is a broader concept that includes emotions and satisfaction over time [12]. If A/B tests only chase short-term gains, it can damage the long-term user experience. Reliable A/B tests should therefore also take into account broader UX aspects, perhaps through guardrail metrics or by being supplemented with qualitative studies. In summary, the path to reliable A/B tests requires a holistic approach where statistical knowledge, methodological rigor, and a sound experimental culture work together.

5 Conclusion

This report has reviewed the most important rules for conducting reliable A/B tests when changing the design in web applications. The review shows that a reliable test is not something you get by chance, but requires thoughtful choices and a careful working method. The report has answered the three questions posed at the beginning, and the answers can be summarized as follows:

The first question was about the theoretical and statistical foundations that are most important for the tests to be reliable. The answer is that you must have:

- **A good theoretical and statistical foundation:** You must understand how experiments, hypotheses, statistical power, p-values, confidence intervals, and effect sizes work. If you misunderstand these things, it is easy to draw the wrong conclusions.

The second question was about the most important practical steps and methods when designing and conducting a test. The answer can be divided into two important rules:

- **A clear strategy for metrics:** Having clear goals and relevant things to measure, including an overall evaluation criterion (OEC) that indicates long-term goals is very important to be able to draw conclusions that mean something and that you can trust.
- **Careful design of the test:** You must carefully calculate how many users are needed, choose a suitable test period, and ensure that you have a good and correct random assignment of users to reduce the risk of errors.

The third and final question was about the most common pitfalls and how to avoid them. To have good protection against incorrect results, you must consider two things:

- **Knowledge of pitfalls and best practices:** You must actively avoid common mistakes, such as testing too many things at once, misinterpreting p-values, or "peeking" at the results. It is very important to follow the advice and rules that exist.
- **Company culture and ethics:** The most important thing of all is the organization's attitude. A culture that appreciates rigor, honest results (even if they are

negative), and that considers ethics is what ultimately creates reliable tests that provide real value.

The most important point is that although A/B testing may seem simple, it requires great knowledge and accuracy to conduct reliable tests. There is a huge 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.