A/B Tests

A/B Test is a quantitative test method to validate or falsify already existing hypotheses. It can be applied to test different versions (A and B) of a website, product, service or prototype by splitting the overall user group into subgroups. The subgroups either only see version A or B, so that the success rate of the alternatives can be measured and compared

NUMBER OF PARTICIPANTS	FACILITATORS	CATEGORY	DURATION	LEVEL OF DIFFICULTY
No Default, Usually Large Samples	Not needed for digital format, in a workshop an instructor is recommended	Test Phase	Multiple hours inclusive analysis of results	Difficult

Description

The method has its origin in marketing: typically in mail marketing, potential customers are often contacted with different mail forms. The individual conversion rates of each of them enables the marketer to find out about the more effective one, that turns more people into customers.
Conversion is referring to a defined process within the customer journey, describing a specific interaction done by the user- ordinarily a purchase, but can also relate to the signing up for a newsletter, the download of a form or clicking on a specific element. Conversion rate is the according KPI that describes the relation of users or visitors of a website to the conversions.
In the example of the direct mail letters, the response rate can provide insight on the factors that influenced the success of the different letters, starting from the appearance (e.g. colour, font size and style) across different contents, wordings or other factors. Nowadays, the method is used in various contexts beyond marketing to measure and compare the conversion or response rate to different concepts or variants.

A/B tests are typically conducted online, without involving any explicit participants or facilitators. They could technically also be done in a workshop, but this is a barely used approach due to the high effort in terms of time and budget. The overall number of users is divided in two groups — A (is typically getting the original version) and B (is receiving the competing alternative to it). In other cases, where there is no original version to be improved, A and B would represent two equivalent alternatives or versions. The splitting of the users in test groups can either happen randomly or dynamically (e.g. based on their profile), whereas the groups eventually need to have the same size.

The test can also be extended to A/B/C/etc. when there are more variants to be compared. Another modification is the so-called multi-variant test, in which the B version differs in more than just one attribute from A. This enables the testing of different combinations, but does not allow measuring the individual effect of the single attributes.

Materials

- In a workshop:
- Two or more prototypes to be tested
- Test Script including the hypothesis to be evaluated and the conversion goals (key metrics to determine the success)
- Spatial separation of the test groups
- Digitally:
- Two or more concepts/designs to be tested
- Hypothesis to be evaluated and conversion goals defined
- Testing Tool to be used, e.g. Google Analytics, Optimizely, Kameleoon, Virtual Website Optimizer

Preparation

As the method applies to the last phase of testing, there need to be two (or more) designs, products or prototypes ready to be tested. It is a quantitative, highly statistical method and therefore requires some statistical preparation. According to the case-specific circumstances, the conductor of the tests needs to define and calculate the appropriate sample size, variance, significance level and "stopping rules" to prevent stopping the tests too early and getting false positive results. A rough benchmark in this case would be a significance level (likelihood of errors that is being accepted) of below 5%. In a simple A/B test, i.e. to compare two different checkboxes, with a baseline conversion rate of 10%, the ideal sample size would be 599.

Step-by-Step Instructions

1. Problem - Define a problem or issue that should be examined in the test.
This can for example be very low click rates on an important element on your website: many users don't sign up for the newsletter.

2. Research — Elaborate on the problem and try to get insights on the causes of the problem, for instance research on design principles and colouring that might be relevant to the website element.
Other website providers have highlighted their newsletter options more prominent. In your web shop, the conversion rates are higher on coloured elements.

3. Hypotheses — Form assumptions about the problem with your research results to have a clear intention for your test. It makes sense to have at least one hypothesis

that can be verified or falsified in your test, so that the test results can easily be interpreted and used for improvement.
A hypothesis for this problem could be: "Many users don't know that we have a newsletter. If the sign-up button was more prominent, more users would perceive it and register."

4. Test — Conduct the test under the pre-set conditions (s. Preparation). Main rules to consider are that the groups need to have the same size, only see the version intended to them- A or B, and that the test should not be stopped before the planned termination.
You implement the B-version containing a highlighted button in a different colour, size or placement and test it against the original version A.

5. Analyse — After the testing phase, the results can be analysed with regards to the hypotheses and the most promising alternative can be determined. If the alternative is the test's "winner", it will replace the original element- if not; you might want to go back to 3. Hypotheses and run another test.
You evaluate the test results after the testing period to find out whether B has caused a higher conversion rate than A.

6. Document — Make sure to document the test set-up and results properly, as it may be useful to review for further improvements in the future or interesting for colleagues or affiliates to understand what has been tested.
Share the results with the team or document it in your knowledge management system. Another colleague might also have issues with the design of important website elements and benefit from it.

Remarks, Tips, Limitations

- For reliable results, the statistical aspect of this method is crucial. There are helpful tools for calculating and defining the right parameters: Sample Size Calculator (Evan's Awesome A/B Tools)
- The right hypotheses for a test are very important to keep track of the goal and end up gaining valuable insights. Take your time to define a clear and meaningful hypothesis: free tools & advice for A/B Testing
- A/B testing works best in the digital environment as described. Nevertheless, it could be conducted in live sessions and workshops by letting a smaller number of users discover the prototypes and observing their handling with it. This allows richer insights and may enable a deeper understanding of the (dis-) advantages of the alternatives. At the same time, the physical testing leads to a loss concerning the objectivity of the results and their statistical relevance.

- A/B testing is a great method to build an empirical foundation instead of gut feeling, as the focus is on the test group, not on the individual, personal view on something. The results can be even more useful if combined with heat or click maps. In general, it can help improve the quality of a website or product feature with reasonable efforts, if the right tools are used.
- A common mistake is to stop the tests too early, which leads to wrong assumptions for results caused by early randomness or no realistic representation of the users (statistical population) yet.
- The method requires big sample sizes to deliver statistically relevant results, so it is rather not applicable for websites/apps with low traffic.

Strengths:

Feasible with reasonable effort if conducted with the right tools. Unambiguous, objective result with clear implications → empirical foundation for decisions. Focus on test group (user/customer base), not individual's perception. Can be extended to A/B/C etc. Strong in combination with heat / click maps.

Weaknesses:

Danger of interpreting data wrong → statistical preparation crucial. Does not work optimally for pages with low traffic or non-digitally. No explanations or follow-ups possible, no insights on the "why".

References

Bedrick, Sarah: "A/B Testing Workshop" Hubspot Academy. https://cdn2.hubspot.net/hub/137828/file-27038161-pdf/docs/ab_testing_workshop.pdf

Breuer, Hendrik (2019): "A/B Testing - Der ultimative Guide mit Expertentipps von Google, HubSpot und anderen" In: Der Shopify Blog. https://www.shopify.de/blog/ab-testing-guide

Liesefeld, Heinrich (2012): "Das Usability-Experiment als Ergänzung zu typischen Usability- und A/B-Tests Inferenzstatistisch abgesicherte Ergebnisse in kleinen Stichproben." In: Usability Professionals 2012 - Tagungsband, Konstanz, Germany, September 9-12, 2012. Accessed through: https://dl.gi.de/bitstream/handle/20.500.12116/5956/Liesefeld_2012.pdf?sequence=2&isAllowed=y

Parikh, Ravi (2014): "Don't Stop Your A/B Tests Partway Through" In: Data Stories. https://heap.io/blog/data-stories/dont-stop-your-ab-tests-part-way-through

Sagebiel, Karl-Fredrik (2020): A/B-Testing: Der ultimative Guide für Einsteiger. https://blog.hubspot.de/marketing/ab-testing

Tullis, Thomas and Albert, William (2008): "Measuring the User Experience. Collecting, Analyzing, and Presenting Usability Metrics." Morgan Kaufmann. Burlington, Massachusetts, USA. Accessed through: https://www.academia.edu/9043043/Measuring_the_user_experience

FAB LAB SIEGEN

fablab-siegen.de

usability-siegen.de