A/B testing

▼ Planning and Prioritizing

1. Set Goals and Define Metrics:

   Example: If your goal is to improve user engagement with a specific feature of your app.

   In this case, you might use metrics such as the number of interactions with that feature, the time spent using it, or the percentage of users who complete a specific action related to the feature.

2. Develop hypothesis:

   Changing xyz feature will increase interactions.

3. Prioritize your tests based on impact and effort

   a. Impact: How much a test will improve your metrics

   b. Effort: Resources required

4. Plan your test design and execution:

   Consider usability testing as well with the A/B test like how is interaction being done on the platform.

5. Analyze your results and learn from them

▼ Purpose of a confidence interval in A/B testing

A confidence interval is a range of values that you think contains the true value of something based on a sample of data. It tells us how sure are we going to be before going into actual scenario. It is a measure of uncertainty around your sample estimate.

Confidence interval helps you measure the uncertainty and variability of your A/B test results. It's a way to show that we're aware our guess might not be certain and that there's a certain level of uncertainty.

The most common statistical tests are z-test and t-test.

Interpreting the results:

> If the interval is positive, that means that your variation is better than your control.
>
> If the interval is negative, that means that your variation is worse than your control.
>
> If the interval is large, that means that there is more uncertainty and variability in your results.
>
> If the interval is small, that means that there is less uncertainty and variability in your results.

▼ P-value

1. What is P-value?

   A p-value is a number between 0 and 1 that represents the probability of observing a difference between two groups as large or larger than the one you actually observed, assuming that there is no real difference between them.

   It's like checking if findings are real or just luck. A low p-value means your A/B test is likely not a coincidence; there's a real impact. It's like strong evidence.

   Researchers typically set a **significance level** (commonly **0.05**), and if the p-value is below this threshold, it indicates that the observed results are statistically significant, supporting the rejection of the null hypothesis.

2. p-values relation to hypothesis testing

   In A/B testing, you typically have a null hypothesis and an alternative hypothesis.

   - The null hypothesis states that there is no difference between the two versions you are testing.
   - The alternative hypothesis states that there is a difference.

- For example, testing a new headline for your landing page,
    - null hypothesis: the headline does not affect the conversion rate.
    - alternative hypothesis: the headline increases the conversion rate.
    - To test your hypotheses, you collect data from your A/B test and calculate a p-value based on the observed difference between the two versions. Then, you compare your p-value to a predefined threshold, called the significance level, which is usually set at 0.05 or 5%

Results interpretation:

- If p-value is less than or equal to the significance level: null hypothesis rejected and alternative hypothesis accepted.
    - This means that you have enough evidence to conclude that there is a difference between the two versions.
- If p-value is greater than the significance level: alternative hypothesis rejected and unable to reject the null hypothesis.
    - This means that you do not have enough evidence to conclude that there is a difference between the two versions.

▼ Power Analysis

A method to find out the sample size that is required to run an experiment.
Variables used:

- Conversion rate
- MDE: explains what is the smallest acceptable difference between the treatment and control groups? It helps understand what minimum effect is worth it, considering the business costs to update
- Statistical significance
- Statistical Power

> 💡 The sample size **increases** if the minimum detectable effect **decreases**.

▼ Randomization

50-50 split does not always mean its a 50-50 split at random. Best case would be to divide the heavy, medium and inactive users equally to conduct the experiment.

▼ Contamination Issues

- Ceteris Paribus: If you want to measure the impact of control vs treatment variant web-page UI, you need to make all other parameters equal i.e. adding a delay factor on control on load time in the control variant as well

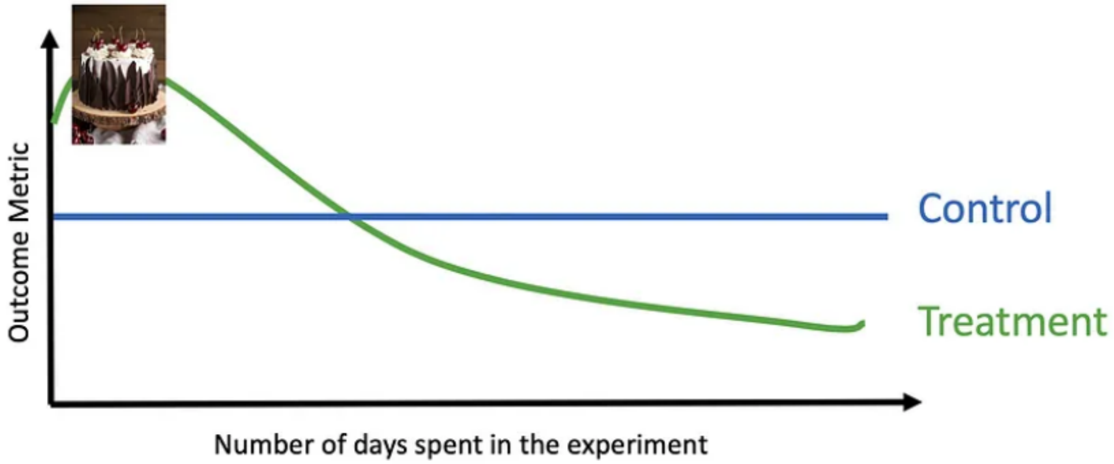
  > 💡 Treatment Group: This group is exposed to the change that is being tested.
  >
  > Control Group: This group does not receive change. Instead, they continue using the existing system without the new feature.

- Spillover effect: The testing was done on city level. Consider Berlin and Munich. If I travel from berlin to Munich, I get exposed to both the usecases leading to a spill over.

▼ Time and its effect in testing

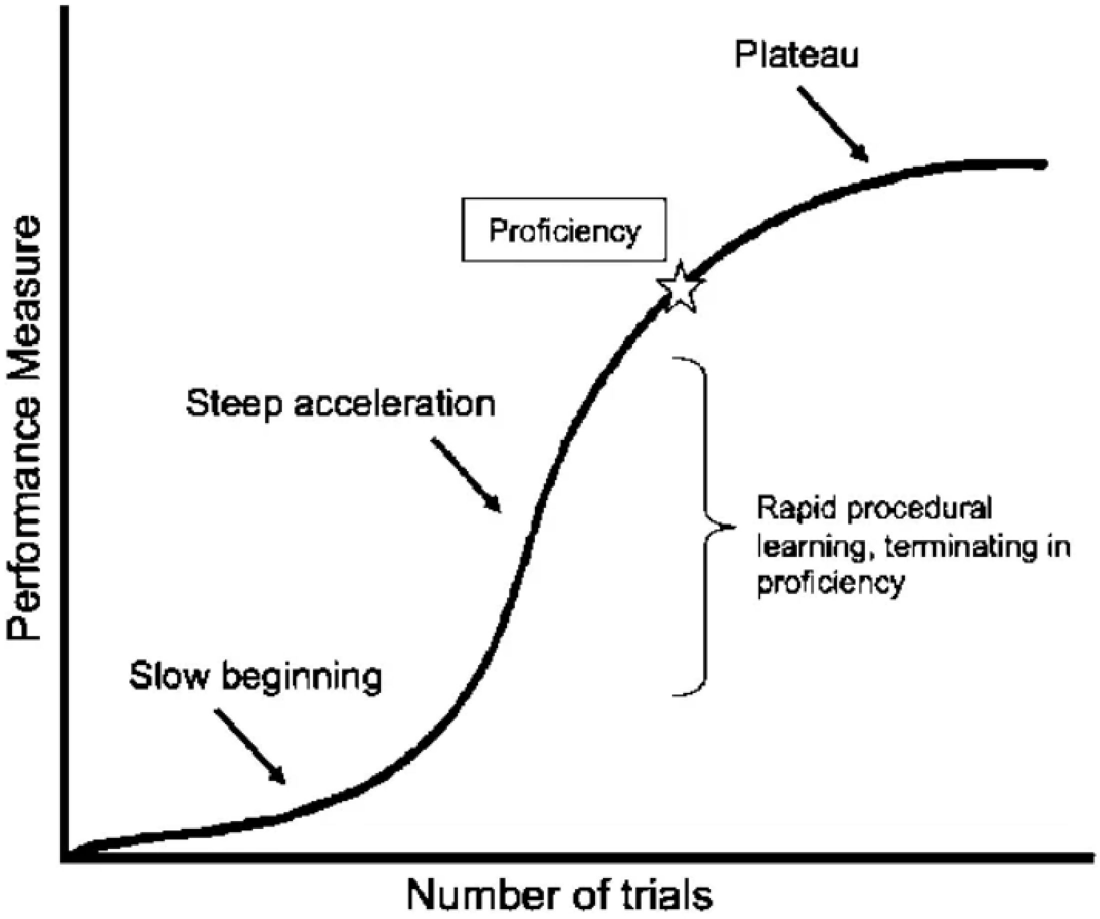
▼ Novelty effect

A novelty effect is a positive effect on a metric because something has changed. Maybe we started adding a cake recipe to the top of our search results. People may interact with our cake recipe a great deal initially, because they are curious, but soon interaction will decrease as people get used to the cake recipe and don’t get value from it.



▼ Learning effect

The inverse of a novelty effect is a learning effect where it takes time for people to become aware of a new feature or learn its value. It takes time for users to trust that the system will be able to accurately respond to a certain kind of input.



▼ Regression to the mean

Regression to the mean is the concept that outliers, positive or negative, will naturally move back to the mean of the population over time.

▼ Types of testing

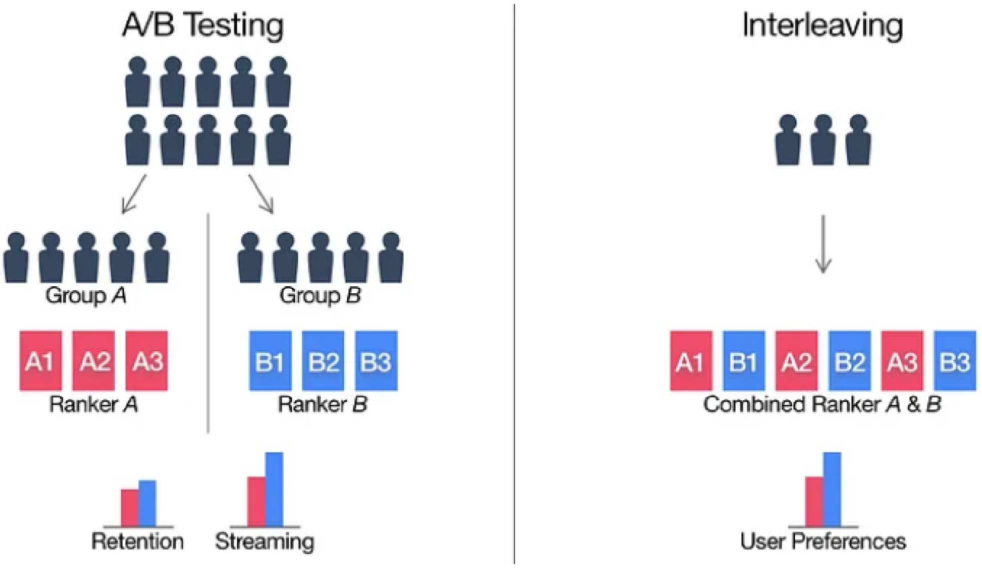
▼ Interleaved testing

Interleaving is an experimental methodology where individual users are shown both variants of an algorithm by interleaving the results.

For example, when a user performs a search, the top results might be a mixed set where results from both A and B are interleaved. The order might be based on a predefined pattern or randomized.

- **Advantages:**

    - Allows both variants to be tested under exactly the same conditions and queries.
    - Immediate comparison of performance as users interact with the mixed results.
- **Disadvantages:**
    - More complex to implement, as it requires dynamic generation of mixed responses.
    - Harder to analyze because user preferences need to be inferred from their interactions with each interleaved result.
- Example: For testing different algorithms that affect the order in which restaurants are displayed after a search query. By interleaving the search results from two different ranking algorithms, Uber Eats can analyze which algorithm leads to more successful transactions.
    - For instance, if Algorithm A places a higher priority on restaurant ratings while Algorithm B emphasizes proximity, interleaving these results for users who search for "sushi" would allow Uber Eats to directly compare the performance of each approach under identical user conditions.



▼ User-Based A/B Testing

User-based A/B testing assigns a specific variant (A or B) to individual users. Once a user is assigned to a variant, they will consistently see the same version each time they interact with the test subject. This method is particularly useful for testing changes that might impact user behavior over time, such as modifications in navigation, user interface, or any feature that benefits from prolonged exposure.

- **Advantages:**
    - Consistency in user experience during the test period.
    - Reduces variability in the test results due to changing experiences.
- **Disadvantages:**
    - Requires tracking and identification mechanisms to ensure the same user sees the same version every time.
- Example: if Uber Eats wants to test a new dashboard layout that highlights gourmet restaurants versus a layout that promotes cheaper quick eats, they might assign half of their user base to each layout. This would allow them to measure which layout leads to more frequent orders over a period of weeks or months.

▼ Session-Based A/B Testing

In session-based A/B testing, the variant is assigned to a session rather than a user. During a session, all interactions are with the same variant, but if the same user returns in a different session, they might be assigned a different variant. This approach is useful when the user's decision is expected to be made within a single visit or session, such as testing checkout processes or single-session user flows.

- **Advantages:**
    - Suitable for tests where user decisions are quick and contained within a single visit.
    - Easier to implement than user-based testing when users are not logged in or identified.
- **Disadvantages:**

  - Potential inconsistency for users visiting multiple times during the testing period.
  - Might not account for changes in user behavior over multiple sessions.
- Example: Suppose Uber Eats wants to test whether a one-page checkout process leads to fewer abandoned carts compared to a multi-page process. By assigning one variant to users for the duration of their session, they can assess which process is more efficient or user-friendly. If a user comes back in a new session, they might see the alternative version, allowing Uber Eats to gather data from the same user on both processes under different sessions.

T-test and Z-test

used to assess the significance of differences between groups.

T-test	Z-test
used when the sample size is relatively small (typically less than 30 observations per group).	used when the sample size is large (usually greater than 30) and the Central Limit Theorem can be applied to assume a normal distribution of the sample.
does not assume that the sample follows a normal distribution, making it suitable for smaller samples with potentially non-normal data distributions.	It is used for larger samples where normality can be assumed.

> 💡 choose the t-test for smaller samples or when normality assumptions cannot be met, and opt for the Z-test for larger samples with normally distributed data.

Resources

Pitfalls in A/B testing