Data Driven Decision making
Best practices for A/B testing - pre-test, test & post-test

1

Define

Based on data, define problems first, and possible solutions second.

Choose ideas to test through a clear **prioritisation model.**

Choose a relevant and measurable Key Performance indicator for your A/B test.

Make sure the measured success (KPI) follows exposure to version A or B, not the other way around. This is also referred to as **temporal precedence.**

Eliminate third variables. If you want to change the text of a CTA, don't change anything else (the location of the button, the color or the size)

Define the **level of measurement** for the independent (the cause) and dependent (the effect) variables

Categorical levels of measurement:

Nominal	Ordinal
The order of items has no meaning	The order of items has a meaning

Numeric levels of measurement:

Interval	Ratio
Intervals are equal, 0 is arbitrary and has no meaning	Intervals are equal, 0 is not arbitrary and has a meaning

Formulate

Formulate a relevant, testable and detailed **test hypothesis** and **null hypothesis** (the opposite of your test hypothesis).

Use the following hypotheses for test where you expect an increase or decrease for a relevant KPI (superiority test)

> **H1: Test hypothesis**
> [Independent variable with intervention] **for** [target group] **on** [page/page group] **leads to** [more/less] [dependent variable] **compared to** [independent variable without intervention]

> **H1: Null hypothesis**
> [Independent variable with intervention] **for** [target group] **on** [page/page group] **does not lead to** [more/less] [dependent variable] **compared to** [independent variable without intervention]

Calculate

Do you expect that your test will result in no decrease/no increase for a KPI? For example; removing content will not lead to a decrease in sales. Use a **non-inferiority approach.**

Define desired **power** (1 - β) and **confidence level** (1 - α) based on risk assessment. Digital Power aims at 80% power, 95% confidence level.

Determine the **minimal detectable effect (MDE)**. This is the minimal effect which, if it truly exists, the test will be able to detect, which justifies implementing the change.

Pre-calculate the **needed traffic/test duration** using minimal detectable effect (MDE), significance level (α), power (1 - β)

2

Decide

Choose a **statistical test** to test your hypothesis based on the collected data, the measurement levels of the variables.

In most cases we are expecting either an increase or decrease in a certain KPI; use a **one-sided test**. Use a two-sided test if you have no expectation about the direction of the effect.

Always check the **assumptions** of your chosen statistical tests and whether your data meets them.

> **P-value < α** **Reject null hypothesis** Assume test hypothesis is true.

> **P-value > α** **Reject null hypothesis** Assume test hypothesis is false.

Test

Testing two groups and proportions.

Check the **equal split of traffic** through a Sample Ratio Mismatch test (Chi Square Test – Goodness of Fit). Found an **SRM error**? Look for the cause to prevent future SRM errors. Filter a specific segment out, or recycle the test.

For tests with two test groups and a proportion as a dependent variable (for example conversion rate, click through rate), choose a **z-test for proportions**.

Beware

Be careful doing deep-dives; the chance of a random significant result (type-1 error) increases with every test. This is also called the **family-wise error rate (FWER)**.

Deep-diving without calculating the needed sample might result in low power; random variation has a larger influence on the results.

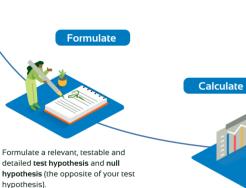**Be mindful running multiple tests at the same time**; it could result in (unintended) interaction effects, especially when they target the same audience, funnel, page and/or KPI.

Conclude

3

Be **clear** on the **conclusion** - if your test hypothesis states an increase in a KPI with no significant result, the conclusion is; it didn't increase the KPI. Not; it decreased the KPI.

Be careful stating **uplift or effect size;** It's largely based on chance and isn't a good predictor for future performance.

Assess

Determine when an outcome of one or more tests is a **learning**.

This can be based on p-value that was found, the number of tests indicating an effect, the number of different locations an effect was found or other indicators (or a combination).

Store

Store your insights or learnings in a database. This way you can share learnings, use them in future tests and combine results of multiple tests.

DIGITAL POWER