



# Running IR Experiments with Real Users - Common Practices and Challenges

Georg Buscher

# A Short Intro

Thesis in IR about measuring attention with eye tracking

2010-2016: building “online” metrics for Bing

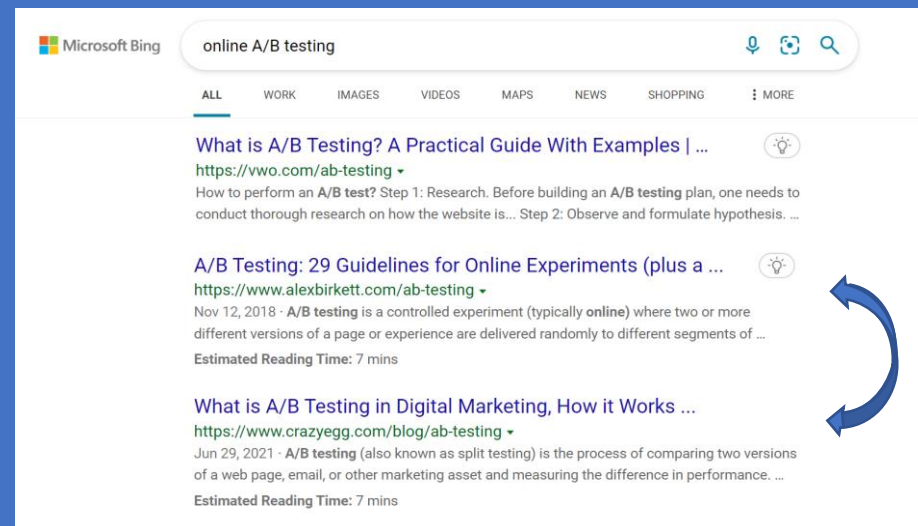
2016-2020: building “online”, “offline”, survey metrics for FB search

2020-2022: experimentation platform in FB/Meta

Since 2022: measurement for Bing

# What Do You Want to Measure?

Search results ranking improvements!



The screenshot shows a Microsoft Bing search page for the query "online A/B testing". The search bar is at the top, and below it are navigation tabs for ALL, WORK, IMAGES, VIDEOS, MAPS, NEWS, SHOPPING, and MORE. The search results are ranked as follows:

- Result 1:** "What is A/B Testing? A Practical Guide With Examples | ..." from <https://vwo.com/ab-testing>. The snippet reads: "How to perform an A/B test? Step 1: Research. Before building an A/B testing plan, one needs to conduct thorough research on how the website is... Step 2: Observe and formulate hypothesis. ..."
- Result 2:** "A/B Testing: 29 Guidelines for Online Experiments (plus a ...)" from <https://www.alexbirckett.com/ab-testing>. The snippet reads: "Nov 12, 2018 · A/B testing is a controlled experiment (typically online) where two or more different versions of a page or experience are delivered randomly to different segments of ... Estimated Reading Time: 7 mins".
- Result 3:** "What is A/B Testing in Digital Marketing, How it Works ..." from <https://www.crazyegg.com/blog/ab-testing>. The snippet reads: "Jun 29, 2021 · A/B testing (also known as split testing) is the process of comparing two versions of a web page, email, or other marketing asset and measuring the difference in performance. ... Estimated Reading Time: 7 mins".

A blue double-headed arrow is positioned to the right of the second and third search results, indicating a comparison or ranking shift between them.

# How Do You Measure?

## DCG!

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)}$$

Result 1  
rel = 5 (**highly relevant**)

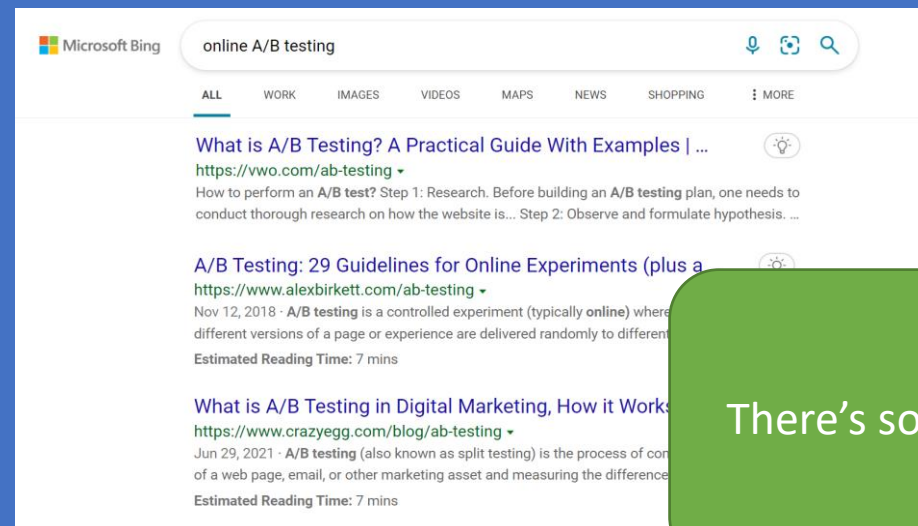
Result 2  
rel = 1 (**junk**)

Result 3  
rel = 4 (**relevant**)

Result 4  
rel = 4 (**relevant**)

What Do You  
Want to  
Measure?

Search results ranking improvements!



There's so much more!

# What Do You Want to Measure?

- Snippets

What is A/B Testing? A Practical Guide With Examples | ... 

<https://vwo.com/ab-testing> ▼

How to perform an **A/B test**? Step 1: Research. Before building an **A/B testing** plan, one needs to conduct thorough research on how the website is... Step 2: Observe and formulate hypothesis. ...

# What Do You Want to Measure?

- Snippets
- Deep links

## Reliable A/B Testing Platform | Trusted by Over 2500 Brands

<https://vwo.com/ab-testing/try-free> ▾ 27K+ Facebook followers

Why: Ad Setup A/B tests in 10 mins with reliable & secure VWO **Testing** platform. Try now! VWO

http: **Testing** offers an unmatched customer support loved globally by 2k+ brands.

How: **Services:** A/B Testing · Multivariate Testing · Split URL Testing

cond

### A/B Testing Guide

Everything You Need to Know About  
A/B **Testing**.

### Try Free for 30 Days

Full featured trial, no credit card  
Onboarding & tech support included

### Split Testing

We offer a wide range of **testing**  
options to meet your needs.

### Contact Us

Get in touch with us today to know  
more about VWO the brand, our

# What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers

The image shows a screenshot of a search engine results page. At the top, there is a blue banner with the text "What Do You Want to Measure?". Below the banner, there is a list of three items: "Snippets", "Deep links", and "Direct answers". To the right of the list, there is a screenshot of a search engine results page. The top part of the screenshot shows a snippet for "Reliable A/B Testing Platform | Trusted by Over 2500 Brands" with the URL "https://vwo.com/ab-testing/try-free" and "27K+ Facebook followers". Below this, there is a dictionary entry for "experimentation". The dictionary entry includes the word "experimentation", its phonetic transcription "[ɪk.sperəmə'n.tʃ(ə)n]", and its part of speech "NOUN". The definition is "the process of performing a scientific procedure, especially in a laboratory, to determine something." There are also two bullet points: "the action or process of trying out new ideas, methods, or activities." and "It was a period of innovation and experimentation with new decorative techniques". A red circle highlights the word "experimentation" in the dictionary entry. Another red circle highlights the word "testing" in the snippet below the dictionary entry.

Reliable A/B Testing Platform | Trusted by Over 2500 Brands  
https://vwo.com/ab-testing/try-free 27K+ Facebook followers

Dictionary Powered by Oxford Languages · Bing Translator

Enter a word Look it up

**experimentation**  
[ɪk.sperəmə'n.tʃ(ə)n]

NOUN

the process of performing a scientific procedure, especially in a laboratory, to determine something.  
"experimentation on the brain and the nerves" · [\[more\]](#)

- the action or process of trying out new ideas, methods, or activities.  
"It was a period of innovation and experimentation with new decorative techniques" · [\[more\]](#)

See more definition

testing



# What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers

The image shows a search engine results page for the query "experimentation".

- Snippet:** "Reliable A/B Testing Platform | Trusted by Over 2500 Brands" with the URL <https://vwo.com/ab-testing/try-free> and "27K+ Facebook followers".
- Dictionary:** A snippet from a dictionary showing the word "experimentation" and a search input field.
- Videos of Experimentation:** A carousel of three video thumbnails:
  - Thumbnail 1:** "What is experimentation?" (1:11) from NESTA - The UK's Innovation Foundation, 2.8K views, Jul 30, 2019.
  - Thumbnail 2:** "Steps in Experimentation | A/B Testing Fundamentals" (3:27) from Data Science Dojo, 3.1K views, Jan 14, 2019.
  - Thumbnail 3:** "'K-Mart of Human Experimentation': Holmesburg Prison's Test" (5:45) from NBC10 Philadelphia, 13K views, Jul 3, 2020.

# What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers
- Utility answers

The image shows a collage of search results for the term 'Experimentation'. At the top, there is a search result for 'Reliable A/B Testing Platform | Trusted by Over 2500 Brands' with the URL <https://vwo.com/ab-testing/try-free> and '27K+ Facebook followers'. Below this is a 'Dictionary' snippet for 'Experimentation' with a search bar and a 'Look it up' button. The main search result is for 'Videos of Experimentation' from <bing.com/videos>. It features a grid of video thumbnails, including one with the text 'A/A TEST' and another with a man's face. Below the videos is a 'Track UPS package' utility with a text input field for 'Enter tracking number' and a 'Track' button. At the bottom, there are three video thumbnails with view counts and dates: '2.8K views · Jul 30, 2019' (YouTube › Nesta - The UK's Inno...), '3.1K views · Jan 14, 2019' (YouTube › Data Science Dojo), and '13K views · Jul 3, 2020' (YouTube › NBC10 Philadelphi). A link 'See more videos of Experimentation' is at the bottom.

# What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers
- Utility answers
- Right/left rail content

The collage illustrates various types of search results and content snippets:

- Reliable A/B Testing Platform**: A snippet with a URL <https://vwo.com/ab-testing/try-free>.
- Dictionary**: A snippet with the text "Dictionary" and a search box "Enter a word".
- Videos of Experience**: A snippet with a URL <bing.com/videos> and a video player thumbnail.
- Track UPS package**: A snippet with a search box "Enter tracking number".
- The Godfather**: A detailed entry for the 1972 film, including a description, ratings (9.2/10 IMDb, 97% Rotten Tomatoes), and metadata (Director: Francis Ford Coppola, Release date: Mar 24, 1972, Gross revenue: \$249.82 million, Produced by: Paramount Pictures, Alfra...).
- Prison's Te...**: A snippet with a video player thumbnail and text "man ion: Prison's Te...".

# What Do You Want to Measure?

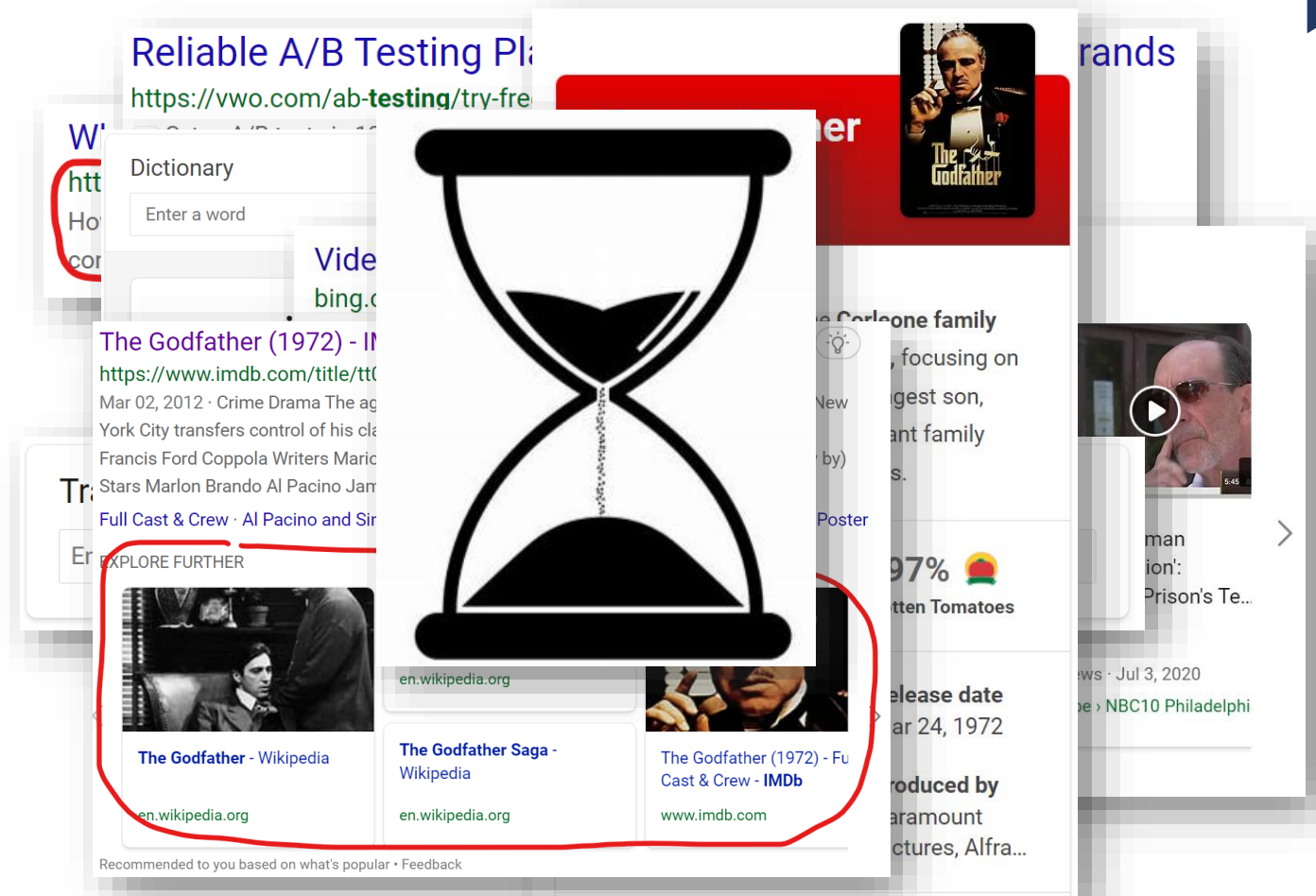
- Snippets
- Deep links
- Direct answers
- Media answers
- Utility answers
- Right/left rail content
- Dynamic content

The image shows a collage of search results for 'The Godfather' film, illustrating various content types:

- Reliable A/B Testing Platform**: A snippet from vwo.com with a URL: <https://vwo.com/ab-testing/try-free>.
- Dictionary**: A snippet with the text 'Enter a word'.
- Videos of Experience**: A snippet from bing.com/videos.
- The Godfather (1972) - IMDb**: A snippet with a URL: <https://www.imdb.com/title/tt0068646>. It includes a description: 'Mar 02, 2012 · Crime Drama The aging patriarch of an organized crime dynasty in postwar New York City transfers control of his clandestine empire to his reluctant youngest son. Director Francis Ford Coppola Writers Mario Puzo (screenplay by) Francis Ford Coppola (screenplay by) Stars Marlon Brando Al Pacino James Caan See production, box office & company info Full Cast & Crew · Al Pacino and Simonetta Stefanelli in The Godfather (1972) · View {Title} Poster'.
- Wikipedia**: A snippet with the text 'EXPLORE FURTHER' and a link to 'The Godfather (film series) - Wikipedia' with URL [en.wikipedia.org](http://en.wikipedia.org).
- The Godfather - Wikipedia**: A snippet with a link to 'The Godfather - Wikipedia' with URL [en.wikipedia.org](http://en.wikipedia.org).
- The Godfather Saga - Wikipedia**: A snippet with a link to 'The Godfather Saga - Wikipedia' with URL [en.wikipedia.org](http://en.wikipedia.org).
- The Godfather (1972) - IMDb**: A snippet with a link to 'The Godfather (1972) - Full Cast & Crew - IMDb' with URL [www.imdb.com](http://www.imdb.com).
- The Godfather**: A red banner with the text 'The Godfather 1972 Film' and a movie poster image.
- 97% Rotten Tomatoes**: A snippet with a Rotten Tomatoes logo and the text '97% Rotten Tomatoes'.
- release date**: A snippet with the text 'release date' and 'ar 24, 1972'.
- produced by**: A snippet with the text 'produced by' and 'aramount ctures, Alfa...'.

# What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers
- Utility answers
- Right/left rail content
- Dynamic content
- Performance
- ...



How Do You  
Measure All  
This?

Online Controlled Experimentation

# How Does It Work?

In a nutshell:

- Split real users into 2 (or more) groups: treatment, and control that get different experiences
- Run the experiment for a while and log whatever you can, importantly: any interactions the user is doing
- Compute metrics based on logged data

What can be re-used?

## DCG-Style Evaluation

- Ground truth!
- Labeled results
- Metric definition



What can be re-used?

### DCG-Style Evaluation

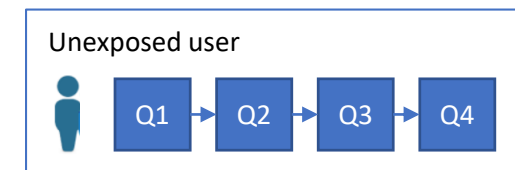
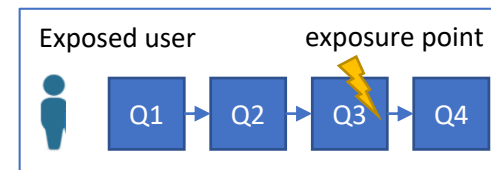
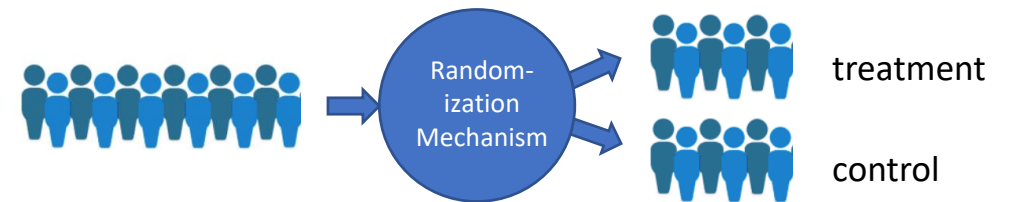
- Ground truth!
- Labeled results
- Metric definition

### Controlled A/B Experimentation

- No ground truth!
- Measurement framework
- Experimentation protocols
- Metrics
- Metric evaluation guidance

# Measurement Framework

- Randomizing users into treatment/control
  - Multiple ways of doing this that can influence chances for false positive metric movements (seed finding!).
- Logging exposure to the intervention
  - Including exposed users only can drastically increase your ability to detect changes.
- Logging system responses
  - Everything from page load time to what results and UI elements were shown to the user.
- Logging user actions
  - Any explicit action taken by the user with timestamp.



```
<Resultspage loadtime="1">  
  <Algo pos="1" title="abc" url="http://abc.de">  
    <Deeplink title="def" url="..."/>  
    <Deeplink title="def" url="..."/>  
  </Algo>  
  ...  
</Result>
```

- Clicks on URLs, dynamic content
- Typing
- Query submission
- Viewport changes
- Mouse cursor movement



# Experimentation Protocols

## Setup

- Declare a hypothesis before starting the experiment
  - What will show the experiment is working as expected?
  - What will show the treatment is better than control?
  - What are important guardrails?
- Size the experiment to properly power the hypothesized metrics

## Start / Monitor

- Figure out appropriate time period to run on
  - Multiples of 1 week to avoid day-of-week effects
  - Long enough so users get over an initial period with novelty effects
- Monitor the health of your experiment
- No accidental changes to the experiment setup while running

## Stop

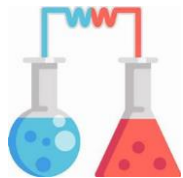
- Double-check metric power to make sure you gathered enough data

# Metrics

## 1. Experiment validity

Does the experiment produce valid data?

- User balance
- Log size
- Invalid log events (e.g., wrong order, etc.)



## 2. System Behavior

Is your IR system behaving as expected?

- Performance
- No results
- Ranking of certain result types
- Client errors



## 3. Behavioral / implicit feedback

Do users react to it favorably?

- Good vs bad clicks and interactions
- Search effort (reformulations, time, ...)
- Abstractions on session and user-day level



## 4. Survey / explicit feedback

Do users like it?

- User satisfaction
- Net promoter score



# Decision Making Guidance

## Check Experiment Validity

- Balanced assignment of users to treatment and control?
- Any metric movements that are surprising could indicate other unfairness

## Abide by hypothesis

- Did the system respond as expected?
- Did users change their behavior as expected?
- Were any guardrails violated?

## Ideal: Back-test experiment

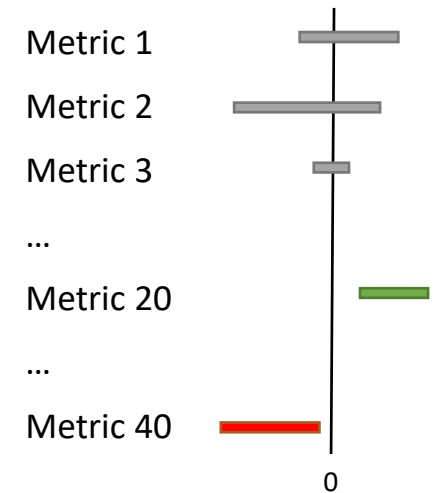
- to confirm gains, particularly when experiment has been of exploratory nature
  - e.g., multiple treatments tested in parallel

# Common Pitfalls and Challenges

Reporting Results from Online Controlled Experiments can Easily Go Wrong!

# Cherry Picking Metrics

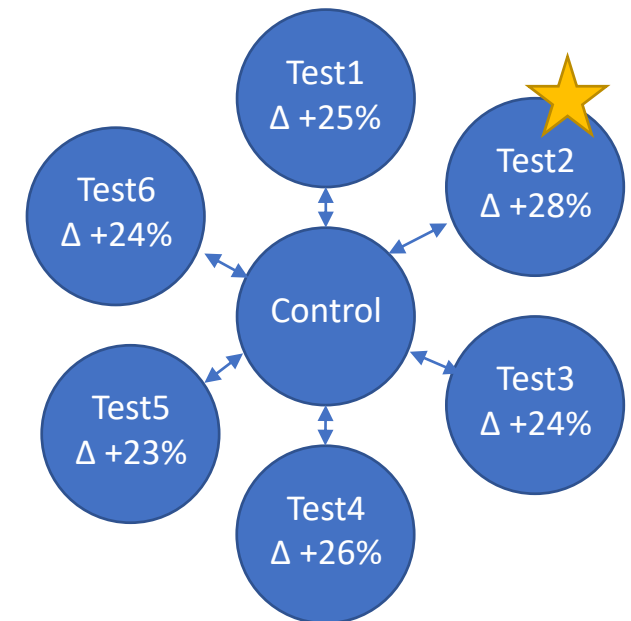
- The more metrics you will look at, the more of them will move.
  - Using Pval of  $< 0.05$  (or 95% conf interval): 1 in 20 metrics expected to move by chance.
  - What is real, what is noise?
- Failure mode: cherry-pick what's green, explain away the red as noise.
  
- Stick to your hypothesis!



# Selecting the Best of Multiple Treatments

- Challenge when running many treatment variants: selection bias
  - E.g., parameter optimization, to find the best parameter setting.
- Selecting the best run from many does not necessarily yield the best setting.
  - Thought experiment: even if you ran multiple parallel A/A tests, one would be the “best”!

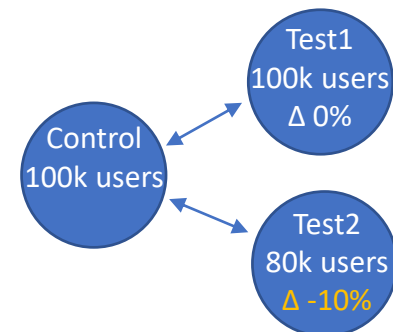
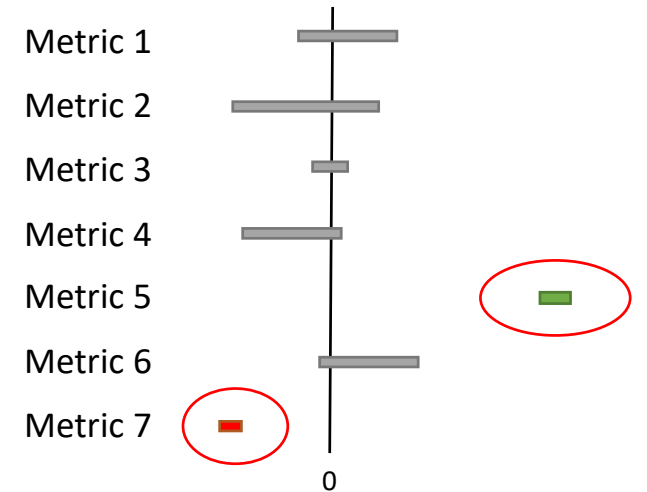
Best practice: Re-run experiment for winning treatment to confirm effect.





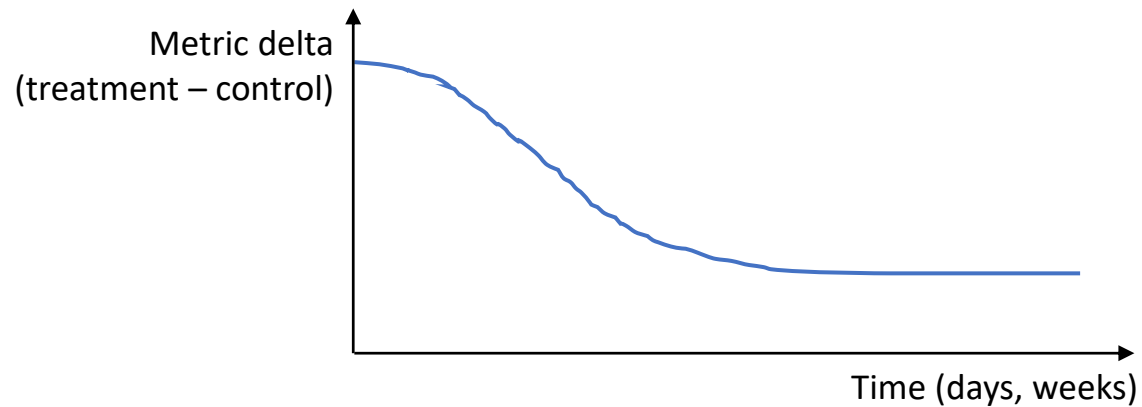
# Surprising Metric Movements

- Surprisingly strong metric movements usually indicate that something is wrong.
  - Never take at face value!
  - Debug and confirm you understand the root cause in all its facets.
- Example:
  - Two treatments, one control
  - Test 2 has fewer users than the others by design
    - Nothing wrong with it per se...
    - But surprising metric movement!
  - → Turns out Test1, Test2, Control each used their own search result cache.
  - → Makes the comparison unfair as their caches warm up at different rates.



# Novelty Effects

- Depending on the nature of the treatment, it may cause a novelty effect.
  - Users react differently for a while (“kick the tires”), then adopt a routine

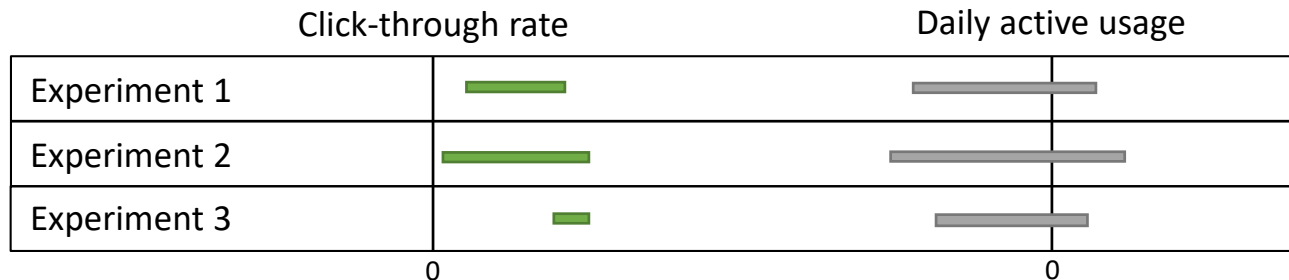


- Awareness for potential novelty effects and explicit investigation whether they exist for a particular experiment are necessary.

# Sensitive Target Metrics vs Insensitive Guardrails

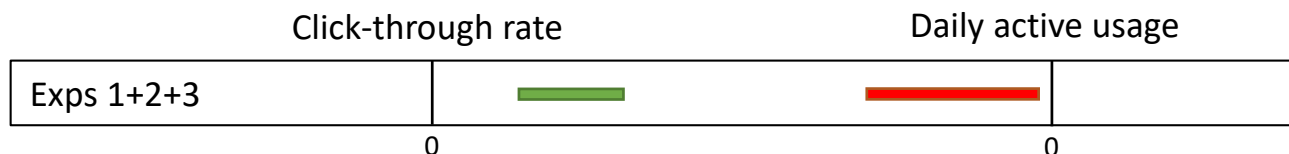
## Example

- Target: increase click-through-rate.
- Guardrail: daily active usage should not regress.



Each experiment increased the target while “not moving” the guardrail. That’s usually enough to declare a win.

- Power analysis really matters!
- Cumulative holdouts can help, if appropriate.



# Summary

## Re-Use in Controlled A/B Experimentation

- Measurement framework
- Experimentation protocols
- Metrics
- Metric evaluation guidance

## Common Pitfalls

- Cherry picking metrics
- Selecting from multiple treatments
- Handling surprising metric movements
- Novelty effects
- Different levels of statistical sensitivity