A/B testing, without fooling yourself.

D8A.ACADEMY/RESOURCES/AB-TESTING-BASICS

BEFORE YOU START

one hypothesis	written down first
one primary metric	decided, not found
MDE	smallest lift you care about

SIZE & DURATION

power = 80%	the convention
n per variant	from MDE, upfront
full weeks only	weekends differ

READING IT

p < 0.05	significant, maybe real
CI of the lift	the honest answer
p = 0.07	not "almost significant"

SPLITTING

randomize by user	not by session
hash(user_id)	stable assignment
A/A test	checks the plumbing

CLASSIC SINS

peeking daily	← manufactures winners.
20 metrics scanned	one wins by luck
stopped at day 3	novelty, not lift

SHIPPING

segment once, after	not to find a win
log the losers too	a test registry
flat is an answer	ship the cheaper one

WORTH MEMORIZING

01 The pre-launch checklist

Everything on this list is decided before the first user is bucketed. That is the whole discipline.

hypothesis	One sentence: we believe X will move Y because Z.
primary metric	One metric. Everything else is supporting evidence.
MDE	The smallest lift worth shipping. Drives the sample size.
duration	Computed from traffic and MDE, rounded to full weeks.
stop rule	The end date. Not the first day p dips under 0.05.

02 Trap: checking daily until it wins

Peek at a flat test 20 times and you get a false winner more than half the time.

YOU DID	WHAT HAPPENS	DO INSTEAD
stopped when p hit 0.049	noise crossed the line first	fixed end date, then read
scanned 20 metrics	one wins by pure luck	one primary, declared
shipped a day-3 spike	novelty effect fades	full weeks, minimum one

→ A 5% false-positive rate is only 5% if you look once, at the end, at one metric.