

 TOOLKIT

The Creative **Testing** System

Test ad ideas on Meta, TikTok and Snapchat, read the numbers against your own target CPA, and scale what sells.



35 pages

10-angle taxonomy

Kill / iterate / scale rules

8-tab planner spreadsheet

Contents

	How to use this toolkit	3
01	Why creative now does the targeting	4
02	Test ideas, not tweaks	6
03	The angle taxonomy	8
04	How many creatives, and how different	10
05	Test structures that work in 2026	12
06	Budget and how long to wait	14
07	The metrics to read at each stage	16
08	Decision rules: kill, iterate or scale	18
09	Fatigue and numbered batches	21
10	Naming that keeps results comparable	23
11	One test across three networks	25
12	The winners library and iterations	28
13	Briefs for the next round	30
	If you want a second pair of eyes	33
	Sources	34

02

Test ideas, not tweaks

A small test can only see a big difference. Different ideas produce big differences. Button colours and font swaps rarely do. That one fact decides what is worth testing.

What a test at your budget can actually detect

Conversions arrive at random. Say an ad's true cost per acquisition (CPA) is exactly \$25. It will not produce one sale for every \$25 spent. Some days it gets three, some days none. So the CPA you see after a few sales is only a rough estimate of the true one. The fewer sales behind it, the rougher the estimate.

The table shows how rough. We calculated it with the standard statistics for counts, an exact 95% Poisson interval. For a given number of conversions, it shows the range the ad's true CPA could plausibly sit in, as a multiple of the CPA you observed.

Conversions behind the number	True CPA could be as low as...	...or as high as
3	0.34× what you see	4.85×
5	0.43×	3.08×
10	0.54×	2.09×
20	0.65×	1.64×
30	0.70×	1.48×
50	0.76×	1.35×

Read the 10-conversion row. An ad showing a \$25 CPA after 10 sales could really be a \$14 ad or a \$52 ad. Say you compare two ads with 10 sales each. A 20% gap between them tells you almost nothing. A 2× gap starts to mean something.

KEY IDEA

At small-business budgets, a creative test can only detect large differences. Design tests whose answers will be large: different ideas, not variations of one idea.

Ideas, executions and tweaks

It helps to name three levels of change.

Level	Example	Typical size of effect	How to test it
Idea (angle)	An offer-led ad vs a customer filming on a phone	Often large: different people respond	As its own cell in a test
Execution	Same angle, with a new hook, new person or new first shot	Medium	As an iteration of a proven angle
Tweak	Button colour, font, one word in the headline, a filter	Usually small; often invisible at your volume	Rarely worth a paid test. Let the platform's automatic options handle it

This does not mean tweaks never matter. At large volume they can. But at 10 to 30 conversions per ad, the noise in the table above is bigger than most tweak effects. A "winning" tweak is usually luck. Spend the test budget where the answer can be seen.

Don't peek your way to a false winner

There is a second trap. If you check a running test every day and stop as soon as one ad looks ahead, you will crown far more false winners than the numbers suggest. Evan Miller's widely used explainer on A/B testing puts it plainly: "Repeated significance testing always increases the rate of false positives". His fix is to "decide on a sample size in advance and wait until the experiment is over" ([Evan Miller, 2010](#)).

The planner builds this in. You set the number of conversions needed to call a result once, on the Settings tab. Before that number is reached, an ad can only be killed for clearly bad numbers. It cannot be crowned early.

WATCH OUT

The most common way to waste a test budget is five versions of one video with different text overlays. Each version gets a fifth of the spend, and none reaches enough conversions to read. To the person scrolling, they are the same ad.

DO THIS

- ☐ List the last 10 ads you launched. Next to each, write its angle in two words.
- ☐ Count the distinct angles. If there are fewer than four, add new angles in your next batch before you add new versions.
- ☐ Take any planned tweak tests (colours, fonts, single words) off the paid test plan.

08

Decision rules: kill, iterate or scale

A rule written down before the test starts protects you from two costly habits: killing good ads too early and keeping weak ads too long. Here are the rules, the reasoning and the probabilities behind them.

The rules

All thresholds are multiples of your target CPA (T), so they work at any price point. The defaults are on the planner's Settings tab, and you can change them there.

Verdict	When	What to do
Too early	Spend below $1 \times T$	Nothing. Read the hook rate and CTR only
Kill: no sale	Zero conversions after $2 \times T$	Switch it off. If the hook rate beat your baseline, save the hook
Kill: early	Spend past $4 \times T$ and CPA above $2 \times T$, before N conversions	Switch it off
Keep running	Between those lines, with fewer than N conversions	No edits. Read again at the next check
Scale	N or more conversions, CPA at or below $1.0 \times T$	Move it to the scaling campaign. Brief 3 iterations
Iterate	N or more conversions, CPA between $1.0 \times$ and $1.5 \times T$	Keep the idea and fix the weak stage (chapter 7)
Kill	N or more conversions, CPA above $1.5 \times T$	Switch it off and log the lesson in one line

Why these numbers

The rules balance two risks: killing an ad that would have worked, and scaling one that won't.

The no-sale kill at $2 \times T$. If an ad's true CPA is exactly on target, the chance it has no sale after spending $2 \times T$ is about 13.5%. At $3 \times T$ the chance falls to about 5%. We use $2 \times$ as the default because a 1-in-7 risk of losing an average ad is cheaper than paying a third time for every loser. If your batches are small and every idea is precious, set it to $3 \times$.

Spend so far (× true CPA)	Chance of zero conversions from an ad that's actually on target
1×	36.8%
1.5×	22.3%
2×	13.5%
3×	5.0%
4×	1.8%

The conversion count to call a result (N). We modelled how often an ad with a given true CPA shows a CPA at or under target once it reaches N conversions.

True CPA of the ad	Looks on-target at N = 5	at N = 10	at N = 20
0.7× target (a real winner)	84%	90%	96%
1.0× target (average)	56%	54%	53%
1.5× target (a loser)	24%	14%	5%
2.0× target (a clear loser)	11%	3%	0.3%

At N = 10, a clear loser (2× target) passes as a winner only 3% of the time, and a real winner is recognised 90% of the time. A middling loser (1.5×) slips through 14% of the time. That's why a Scale verdict leads to a first week of watching in the scaling campaign, not to a big budget overnight. If your volume allows N = 20, the error rate on middling losers drops to 5%.

KEY IDEA

The rules don't make you right every time. They make you wrong in small, cheap ways instead of large, expensive ones, and they make the same decision every time, whoever is reading the numbers.

Reading the Next move column

For Iterate and Kill verdicts, the planner suggests where to look, using your baseline hook rate and CTR:

- **Hook rate below baseline** → keep the body, test 3 new hooks.
- **Hook fine, CTR below baseline** → the opening works but the middle or the call to action doesn't. Rework those.
- **Hook and CTR fine, CPA high** → people click but don't buy. Check that the offer and landing page match the promise in the ad.
- **Killed, but hook rate above baseline** → the opening stopped people. Save it in the Winners library notes and reuse it on a stronger idea.

EXAMPLE

These rows come from the planner's sample data, at a \$25 target CPA.

Test ID	Spend	Conv.	CPA	Hook rate	Verdict	Next move
B01-02	\$460	23	\$20.00	31.0%	Scale	Move to scaling, brief iterations
B01-03	\$440	14	\$31.43	26.8%	Iterate	CTR fine: check offer and landing page
B01-04	\$300	3	\$100.00	18.0%	Kill (early)	Log the lesson
B01-08	\$55	0	–	20.0%	Kill (no sale)	Log the lesson
B01-12	\$20	0	–	21.4%	Too early	Leave it alone

VERSION AFTER VERSION, UNTIL TWO ADS CARRIED THE STORE

For an incense charcoal brand, we rebuilt the sales campaigns in versions. Each new version kept the winners from the last and replaced the rest: one ad set, several ads, clear names, so every result could be traced to the creative behind it. Two ads ended up driving about 95% of all purchases. One alone brought 810 orders at a 5.72 return. Over two months, the account reached 1,651 orders at a 5.07 return. [Read the project](#)

DO THIS

- ☐ Type your target CPA into the planner's Settings tab. Leave the other rules at their defaults for your first two batches.
- ☐ Choose your check days (for example, Monday and Thursday) and only read verdicts on those days.
- ☐ After two batches, review the kills. If you regret more than one, raise the no-sale kill to 3×.



AdvanSEM

— ADVANCE YOUR MARKETING —

Advance your marketing.

advansem.com · info@advansem.com