INTERNAL BRIEF • MEASUREMENT

Incrementality Testing

What it is, what we still need to learn, and how we identify and build our first credible test

---

Companion to: Performance Optimization Framework : Internal Project Brief (v2)
Audience: Media Activation Optimization Team + Partner Success
Purpose: Educate the team, identify the right first pilot, and build a test we can run, learn from, and eventually repeat across clients.

---

1. Why we need this

Our Performance Optimization Framework already makes the tension clear: we should not move budget to whichever channel has the highest last-click ROAS, but we also should not protect upper-funnel investment just because attribution cannot fully measure it. Today, the signals we use to understand that contribution, including view-through activity, downstream behavior, and the Display-to-Search halo, are directional. They help us make a smarter recommendation, but they are still inference rather than causal measurement.

Incrementality testing gives us a way to move from that directional evidence to a controlled estimate of impact. It asks a different question than attribution: what changed because the media was present, compared with what likely would have happened without it?

> The core idea in one line
> Incrementality estimates what changed because the media was present compared with what likely would have happened without it. Attribution assigns credit; incrementality tests contribution.

2. What it actually is

Attribution tells us which channel received credit for a conversion. Incrementality asks whether the outcome would have happened without that media exposure in the first place. The basic idea is to create a credible comparison between a group that receives the media and a comparable group that does not, then measure the difference in outcomes.

- **Exposed group:** a comparable set of people (or geographies) that see the channel's ads.
- **Holdout group:** a matched set that sees no ads for that channel : everything else held identical.
- Lift: the difference in outcomes between the two groups, after accounting for the test design. That gap is our estimate of incremental contribution.

> Worked example
> If a comparable exposed group generates 1,000 ticket sales and the holdout generates 850 over the same test window, the observed difference is 150 sales, or 15%. In a real test, we would still need to account for group size, baseline differences, and statistical uncertainty before calling that the incremental effect.

3. The three ways to run one

There is not one universal incrementality test. The right method depends on the business question, the channel, the data we can access, and how cleanly we can create a control group.

Geo holdout

Split markets into comparable geographic groups. The test geos receive the media being evaluated and the control geos do not. When we have reliable geo-level ticket sales or revenue, this can connect the experiment directly to the business outcome, which makes it a strong candidate for our first arts pilot.

Trade-off: needs enough geos and enough conversion volume to reach significance, and you are deliberately withholding spend from real markets for the test window.

Platform-native conversion lift

Platforms such as Meta, Google, and TikTok offer native lift studies that create randomized exposed and control groups within their environments. These can be a practical way to measure incremental conversions or brand outcomes when the platform, account, and campaign meet the eligibility requirements.

Trade-off: these studies are platform-specific, may have eligibility or minimum-volume requirements, and usually answer a narrower question than a cross-market business-outcome test. They are useful tools, but the method should be selected based on the question we need to answer.

Ad-level or randomized holdout

Where the buying platform and setup allow it, users or impressions can be randomized into treatment and control conditions so the media being tested is the primary difference between groups. This can provide clean isolation, but the exact implementation and outcome measurement depend on platform capabilities and should be validated before launch.

Method	Best for	Measures	Rigor	Effort
Geo holdout	Display, CTV, TikTok, Search; multi-market clients	Business outcome when geo-level sales/revenue is available	High	High
Platform-native lift	Meta, Google/YouTube, TikTok where eligible	Platform-defined conversion or brand outcomes	Med-High	Low-Med
Randomized/ad-level holdout	Channels/platforms that support a valid control design	Depends on setup and measurement source	High when well designed	Med-High

4. The first decision: test one question well

> **Do not start by testing the entire mix at once.**
> **The temptation is to prove all six channels in one test. That may show whether the bundle mattered, but it will not tell us which channel drove the difference or what to do next.**
> **For the first pilot, answer one question well: does Display create incremental ticket sales beyond what would have occurred without Display? Search and other downstream behavior are the second read.**

5. The pilot test we should run

Our first pilot should turn one of our most important directional assumptions into a measured result. The leading question is whether Display creates incremental ticket sales beyond what would have occurred without Display. Search is an important second read because it can help us understand how that contribution may be showing up, but it is not the outcome we are trying to prove.

Element	Specification
Hypothesis	Display (DV360) creates incremental ticket sales beyond what would have occurred without Display.
Method	Leading design: geo holdout with comparable test and control markets. Display runs in test geos and is withheld in control geos; other media and major business conditions are kept as comparable as possible. Final design is confirmed after client qualification and power analysis.
First read: business outcome	Total ticket sales or revenue by geo when available. This is the primary measurement outcome and answers whether Display created incremental business value. If geo-level ticket sales are not available, define the strongest reliable business outcome before launch.
Second read: downstream behavior	Search conversion volume, branded search demand, site traffic/engagement, Meta conversions, and other downstream signals available for the client. These do not determine whether Display was incremental; they help us understand where and how the measured lift may be showing up.
Duration	Set by the power analysis and a clean business window, ideally within a stable sales phase. Do not default to a fixed 4 to 6 weeks if the volume or sales cycle requires something different.
Read	First, estimate the incremental difference in ticket sales/revenue between test and control and report the uncertainty. Then use the second-read signals to understand the likely pathway of that lift. A null or non-significant first read must be interpreted against the detectable effect size and test quality before concluding there is no value.

6. How we would build it

This is the working build process. The first steps are intentionally qualification and feasibility, because we should not commit to a client or test design until we know the question can be measured credibly.

Step 1: Qualify the client and pick the question

- Confirm the client has enough investment and outcome volume in the channel relationship we want to test, and that answering the question would change a real media decision.
- Confirm the outcome data we can access at the level required by the design, ideally ticket sales or revenue by geography for an arts pilot, with site conversions as a possible fallback depending on the question.

Step 2: Run a power analysis before we commit

- Estimate whether the available volume, number of markets, expected effect size, and test window give us a reasonable chance of detecting a meaningful lift. We should define the minimum effect

worth detecting based on the client and decision, rather than assume 5 to 10% is always the right threshold.

- If the test is unlikely to detect a business-relevant effect, redesign it before launch. That may mean a longer window, a different client, larger geographic cells, or a different channel question. An inconclusive result from an underpowered test should not be interpreted as proof that the channel has no impact.

Step 3: Build comparable test and control groups

- Use the pre-period to create groups that are as comparable as possible on the outcomes and market characteristics that matter. Depending on the number of markets and test design, this may use matched pairs, stratification, or another experimental design rather than one fixed matching approach.
- Check that the groups behave similarly before the test. If baseline outcomes are materially different or trends do not move together, revisit the design before launch.
- Reserve enough geos on each side. Too few and one unusual market swings the whole result.

Step 4: Lock the test design and control other changes

- For a Display geo holdout, Display would be withheld in the control geos while the rest of the media strategy is kept as comparable as possible across groups. Any differences that cannot be held constant need to be documented and accounted for in the readout.
- Avoid major creative, budget, targeting, promotional, or channel changes during the test window when possible. Maintain a change log for anything that does move.
- Watch for contamination: audiences that cross geo lines, retargeting pools that carry control users into exposed inventory, and national buys that ignore geo. Suppress or exclude these.

Step 5: Run the test and protect the design

- Run for the planned window unless there is a delivery, tracking, or business issue that requires intervention. Do not optimize the treatment differently because of early performance reads; that changes the question the test is answering.
- Keep the change log current so any anomaly can be explained later.

Step 6: Measure lift and uncertainty

- Estimate the difference in outcomes between treatment and control using the analysis method defined before launch. Report the estimated lift together with its uncertainty or confidence interval, not just a single point estimate.
- State clearly whether the result is statistically distinguishable from zero and whether the test was powered to detect the effect we care about. A non-significant result may mean there is little incremental impact, that the effect is smaller than our detectable threshold, or that the test is inconclusive. The readout should distinguish those possibilities.

Step 7: Translate the result into a media decision

- Bring the result back into the Optimization Framework and the Next-Dollar Test. Does the measured contribution support keeping or scaling the investment, changing the tactic, running another test, or moving budget elsewhere?

- Bank the test design, result, incremental ROAS or cost per incremental sale when appropriate, and the conditions under which the result was observed. One test becomes a case study; repeated tests are what eventually become a useful benchmark library.

7. What can make the test unusable

The biggest risks to plan for:

1. Weak test/control comparability. If the groups are materially different before launch, the post-test gap is difficult to interpret. Spend time validating the pre-period.
2. Insufficient volume. If the test cannot detect a business-relevant effect, a null result may be inconclusive rather than evidence of no impact. Run the power analysis first.
3. Contamination. Control audiences seeing the test media, cross-geo delivery, retargeting leakage, or national buys can reduce the difference between treatment and control.
4. Major outside changes during the test. Sales-phase shifts, promotions, pricing changes, weather, local events, or other media differences can affect outcomes. We should choose the cleanest window possible and document what we cannot hold constant.

8. How we turn this into a repeatable capability

The first test is how we learn the process. If we can run it credibly, the next step is to standardize what should be repeatable without pretending every client or channel needs the exact same design:

- A test brief template: business question, hypothesis, method, primary outcome, test/control definition, power analysis, duration, success criteria, and known risks, completed before launch.
- A pre-launch checklist: feasibility confirmed, groups validated, measurement plan locked, contamination risks addressed, client constraints documented, and change log open.
- A standard readout: estimated lift, uncertainty/confidence, what the result does and does not prove, media decision, and next learning opportunity.
- A benchmark library: test results by channel, client type, sales phase, and test conditions. We should only generalize once we have enough comparable evidence to support it.
- A test roadmap: sequence the questions with the highest decision value. Start with whether Display creates incremental ticket sales, using Search and other downstream behavior as the second read. From there, move to other high-value questions such as CTV-to-total-sales once we have a client and design that can measure them credibly.

> The client story we are working toward
> “We want to move beyond attribution alone and test what our media is actually adding. Controlled experiments give us a stronger basis for deciding where the next dollar should go, and where we need more evidence before making a change.”