guide

2026

measuring true advertising impact

a guide to incrementality experiments

iab. australia

about this paper

This whitepaper has been developed by the **IAB Australia Advertising Effectiveness Council.** The Council brings together representatives from media owners, agencies, advertisers, measurement and research companies and technology vendors. The Council develops industry guidance and resources to support better approaches to measuring the impact of marketing activity and help marketers make more informed decisions about their advertising investment.

This paper aims to build industry understanding of incrementality, causation, the principles of rigorous experimentation, and the conditions required to produce reliable and actionable ad effectiveness insights.

Controlled experiments provide a rigorous way to establish the causal impact of a marketing activity or strategic change on a defined outcome. When designed and executed well, they can provide a strong evidence base for validating existing activity, addressing knowledge gaps and informing optimisation decisions. The paper sets out the standard of best practice required for experiments to provide a reliable source of truth, while recognising that experiments are one part of a broader measurement approach.

By reading this paper, marketers will:

- Understand what good experimentation looks like and the key principles that underpin a rigorous advertising effectiveness experiment.
- Know how to assess and plan an experiment, including how to define incrementality, establish robust test conditions and avoid common pitfalls across different outcomes and digital media channels.
- Understand how to use experimental results with confidence as part of a broader measurement approach, alongside techniques such as MMM and attribution.

> *"Marketers are being asked to show that their advertising does more than take credit for sales that would have happened anyway, and incrementality measurement is the most direct way to answer that. Yet different incrementality methods vary widely in rigour, and finding the best fit carries a number of crucial considerations. This paper was written to help with that choice. It is a collaborative work by some of Australia's leading minds in advertising effectiveness, pairing real depth of expertise with guidance marketers can use straight away. We hope it gives them a stronger footing for incrementality discussions and more value from the solutions they rely on."*
>
> **— Dave Goodfellow, Head of Marketing Science at Tracksuit and Chair of IAB Australia Ad Effectiveness Council**

IAB Australia Advertising Effectiveness Council would like to sincerely thank the following companies for their contribution to content: Amazon Ads, Billups, Coles360, dentsu, Google, Kantar, Lifesight, Mars Petcare, Meta, Monks, Netflix, Prophet, RyvalMedia, Samba, Seven Media Group, Tracksuit.

The contents of this guidance paper are not necessarily reflective of individual company policies, rather it is a collaborative effort of IAB Australia councils and working groups to provide greater understanding to the market.

executive summary

Marketers are under growing pressure to show that every advertising dollar is doing real work. As privacy changes, signal loss and platform restrictions make user-level attribution more challenging, incrementality experiments offer a rigorous way to distinguish the additional outcomes driven by advertising from those that would have happened anyway.

But a lift figure is only as credible as the experiment behind it. The quality of the design, the reliability of the data and the way results are interpreted all determine whether the evidence can support a confident marketing decision.

This summary distils the principles of rigorous experimentation into ten things every marketer should know before commissioning, running or acting on an ad experiment.

getting the fundamentals right

1. **Incrementality is about causation.** Experiments estimate the impact of advertising beyond what would have happened anyway.
2. **A credible causal result needs three things:** a valid counterfactual (what would have happened without the advertising), control of other factors that could influence the outcome, and enough reliable data to distinguish a real effect from chance.
3. **Not all methods provide the same strength of evidence.** Incrementality is not a single technique but a family of approaches. What unites them is the goal of isolating the true business impact of marketing activity. But not all methods are equal in their ability to establish causality. Randomised experiments, holdouts and matched-market tests are strongest. Model-based approaches help where randomisation isn't possible but rely on more assumptions.
4. **Use experiments to answer the right business question.** Experiments can be used to validate brand and sales impact, compare creative and tactics, inform decisions to scale or stop, and calibrate other models. Market Mix Modelling is better suited to setting overall budgets and channel allocation.

designing a test that can be trusted

5. **Start with the decision, then design the test.** Define the hypothesis, the outcome and the decision the result will inform, and calculate the minimum detectable effect and required statistical power before launch. Ensure the ad creative links to the business outcome. Draw on previous results, understand what information is already available and what have you learnt from the past.
6. **Keep conditions consistent and spend at a meaningful level.** Spend must be enough to detect an effect without pushing the channel into saturation, and results at one spend level should not be assumed to apply at another.
7. **Measure over a period of time that reflects the consideration window.** Define it upfront, including whether a post-campaign period is needed to capture the full effect.
8. **Apply the same principles in every channel.** The considerations below apply to every experiment, but shared screens, high-intent audiences, modelled exposure and fragmented media change how they are met. Exposure classification should use the most accurate method available for the channel.

key considerations that apply across all media channels

- **Credible counterfactual:** establishing what would have happened without the advertising.
- **Randomisation or appropriate control construction:** ensuring treatment and control are comparable and minimising selection bias.
- **Control contamination:** preventing the control group from receiving the treatment or otherwise being affected by it.
- **Accurate exposure measurement:** confirming who received the treatment and at what level.
- **Sufficient sample size and duration:** ensuring the experiment has enough statistical power to detect an effect.
- **Control of other influences:** accounting for factors such as seasonality, other marketing activity and external events.
- **Appropriate outcome and measurement window:** allowing sufficient time for the effect to occur and defining the outcome consistently.
- **Validation and transparency:** documenting methodology, assumptions, data quality, limitations and uncertainty.

- **interpreting and acting on results**

9. **A significant result is not necessarily commercially valuable, and an inconclusive one is not proof of zero impact.** Judge the size and uncertainty of the effect against the business's economic threshold. Results apply to the conditions tested.

10. **Experiments are the causal anchor of a broader measurement approach.** MMM supports strategic allocation and attribution supports optimisation; experiments validate and calibrate both. Different methods often produce different answers, this does not necessarily indicate that one method is wrong and the reasons for the difference should be understood.

the bottom line

Incrementality experiments help marketers understand what advertising actually adds to business outcomes—beyond what would have happened anyway. But the value of an experiment depends on the strength of its design, the quality of its data and the transparency of its methodology. A reported lift alone is not proof of impact.

When conducted well, experiments provide robust causal evidence to validate advertising effectiveness, guide investment and optimisation decisions, and calibrate MMM and attribution. Used together, these approaches give marketers a more complete understanding of advertising's contribution and a stronger evidence base for deciding where investment will deliver the greatest value.

contents

1 Why marketers need to measure incremental impact 6

2 What incrementality means and what makes a measure causal 8
- The key elements of causal measurement 8
- What best practice looks like 9
- The causal-strength spectrum 10
- What incrementality is not 11

3 When experiments are useful: strengths, limitations and use cases 12
- Strengths 15
- Limitations 16

4 Choosing an approach: experiment types and how they establish incrementality 17
- Experiment-based (strong causal strength) 17
- Model-based counterfactual (strong to moderate causal strength) 18
- Econometric (moderate to weak causal strength) 19
- Non-causal methods (don't measure incrementality) 19

5 Designing a reliable experiment 20
- Laying the groundwork before the test 20
- Establishing a reliable baseline 20
- Setting the intervention period 21
- Setting the right level of media spend 21
- Measuring the post-intervention effect 21
- Considerations across purchase cycles 22

6 Considerations for measuring incrementality by channel 24
- Social media 25
- Paid search 25
- CTV/BVOD/streaming 25
- Retail media 27
- Digital OOH 28
- Streaming audio 29
- Podcasts 30
- Creator / influencer 30

7 Putting experiments in context: how they work in the measurement toolkit 32
- Why results across MMM, attribution and experiments may not align 33
- Using brand lift experiments and sales lift experiments in tandem 34
- Using experiments to calibrate attribution and MMM 35

8 Note on changes to privacy legislation 40

9 Transparency template for incrementality experiments 41

A Appendix — more definitions and terms 45

B Appendix — find out more 46

1 why marketers need to measure incremental impact

↑ back to contents page

Incrementality—the additional value generated by marketing investment beyond what would have happened without it—is a critical measure of advertising effectiveness. It provides a way to distinguish the outcomes caused by advertising from those that would have occurred anyway, helping marketers understand the true contribution of their marketing activity.

Randomised controlled experiments provide one of the most robust approaches to measuring incrementality because they can establish a direct causal relationship between a marketing intervention and a defined business outcome. By creating comparable test and control groups, experiments help isolate the impact of a strategic change and provide greater confidence that observed differences are attributable to the marketing activity being tested.

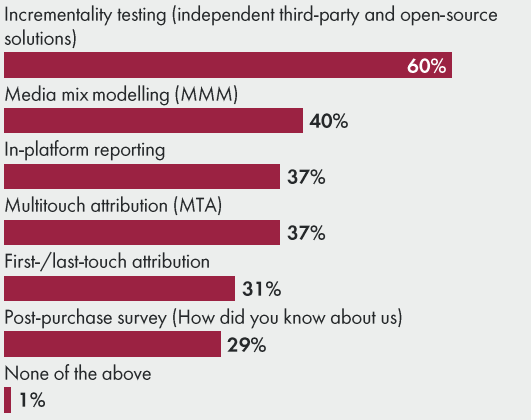
When designed and conducted to a consistent standard, experiments can provide a rigorous evidence base for validating existing marketing activity and informing investment and optimisation decisions. They can be used to measure a range of outcomes, from brand effects through to sales and other business actions.

Marketers have access to a growing range of approaches for estimating the incremental impact of advertising, and **the accuracy and reliability of these approaches can vary depending on the methodology, data, experimental design and conditions in which they are applied**. The quality of the evidence is therefore critical to the confidence placed in the resulting decisions.

Marketers are placing greater weight on incrementality as the need for more reliable evidence of advertising impact grows. The decline of third-party cookies, privacy-driven changes to tracking and increasing platform restrictions have made user-level attribution more challenging, while economic uncertainty has increased pressure to demonstrate the value of every marketing dollar. Incrementality experiments provide a way to isolate the additional impact generated by advertising beyond what would have occurred organically.

Most Trusted Marketing Measurement Solutions Among US Senior Decision-Makers, Jan 2026

% of respondents



Source: Haus, "2026 Marketing Decision Confidence Index," Mar 18, '26

EMARKETER

This growing emphasis is reflected in the Haus 2026 Marketing Decision Confidence Index, reported by eMarketer, which found that **60% of US senior decision-makers said they trust independent incrementality testing most** among measurement solutions—20 percentage points ahead of MMM (40%) and substantially higher than in-platform reporting (37%).

The IAB US *State of Data Report 2026* reported that incrementality tests are the most common advanced measurement approach used (76% of US buy-side decision makers use incrementality experiments, 73% attribution analysis and 67% Market Mix Models). Yet only 39% use all three together, despite each serving a distinct purpose, leaving most marketers without the holistic view of ROI that today's landscape demands. Robust incrementality experiments play an important role in the measurement mix, establishing the causal impact of advertising by measuring the additional outcomes it generates beyond what would have happened without it. Marketers should have the goal of working with all their chosen vendors and providers to ensure incrementality testing meet appropriate standards of rigour and transparency.

Each methodology has a role to play. IAB Australia's *Future of Measurement Report 2026* found strong industry support for integrating complementary measurement methods within a common effectiveness framework. Rather than treating methodologies as competing alternatives, this approach recognises that different methods are suited to answering different business questions.

An emerging effectiveness framework brings together:

- Market Mix Modelling (MMM) for strategic budget allocation and understanding total marketing impact.
- Attribution for operational optimisation where appropriate.
- Incrementality experiments for causal validation and calibration.
- Brand and quality metrics for understanding longer-term outcomes and media value.
- Headline business performance metrics for tracking overall business outcomes and providing context for interpreting measurement findings.

This direction is consistent with IAB US's *Modernizing MMM guidance*, which advocates cross-validation and the use of multiple supporting signals when making significant marketing investment decisions.

The opportunity for marketers is to understand what each approach can reliably tell them, apply it to the business questions it is best equipped to answer, and combine the resulting evidence to build greater confidence in marketing decisions (more detail on combining and calibrating other methods with incrementality experiments is provided in section 7).

This paper focuses on **how to design and conduct rigorous advertising effectiveness experiments, with particular emphasis on incrementality and causation**, so that experimental results can provide reliable test-and-learn insight and complement other measurement approaches within a broader effectiveness framework. While 'incrementality' is sometimes used as a broad term for measuring advertising impact, demonstrating a true causal effect requires specific conditions and methodological practices (as outlined in section 2).

2 what incrementality means and what makes a measure causal

↑ back to contents page

IAB Australia has adopted the IAB US and IAB Europe definition of incrementality and the framework for primary incrementality methods outlined in IAB *Guidelines for Incremental Measurement in Commerce Media*:

Incrementality measures the causal impact of marketing by identifying the additional business outcomes directly driven by a campaign or tactic, compared to what would have occurred in the absence of marketing activity.

Essentially answering the question: What changed because of the marketing action, versus what would have happened without it?

the key elements of causal measurement

For a method to credibly claim causality, the IAB / IAB Europe framework sets out three requirements — the key elements of causal measurement.

- **A credible counterfactual or intervention.** Establish a valid "what if not" scenario or a clearly defined intervention that separates test and control groups. The counterfactual represents what outcomes would have looked like in the absence of the marketing action. Interventions create the necessary conditions for comparing when treated versus untreated groups (e.g., user-level or geo-level separation).
    - **Why it matters:** Without a credible counterfactual, measured differences may reflect correlation rather than true campaign impact. The rigour of this initial design is the primary factor determining the causal strength of the entire model.
- **Controlling for bias.** Ensure the measured effect reflects the campaign's impact rather than other hidden factors that influence the outcome. In practice, it is rarely possible to remove all sources of bias, but well-designed causal methods can reduce it to a level where results are still actionable and useful.
    - **Why it matters:** Bias can easily distort results, leading to misallocation of budget. Factors like seasonality, pricing, promotions, competitor activity or unmeasured concurrent campaigns can make results misleading if not accounted for (omitted variable bias).
- **Separating signal from noise.** Distinguish true incremental effects from randomness, sampling variability, or insufficient data. Noise can never be fully removed, but well-designed studies can reduce its influence and quantify the uncertainty around results.
    - **Why it matters:** Even if the design is conceptually sound, a lift estimate that is statistically insignificant (e.g. falls within the noise floor) is not reliable for decision making. A failure to detect a real effect due to small sample size represents inefficient allocation of testing budget.

what best practice looks like

The **Media Rating Council (MRC)** standard outlined in their Outcomes and Data Quality Standards 2022, describes incrementality in the same terms: the measure of the true value created by any business strategy, determined by isolating and measuring related results, independent of other potential business factors. In other words, incrementality refers to the causal impact of marketing.

Approaches include randomised control experiments, also known as Randomised Control Tests (RCTs), and non-experimental (or observational) methods. RCTs and non-experimental methods differ in how test and control groups are created.

In summary, the MRC minimum standards for test-and-control experiments include:

- **Use randomisation for true experiments** – In an RCT, individuals must be randomly assigned to test and control groups, with the researcher controlling who receives the treatment. Any approach that does not control the treatment of individuals and/or does not do this at random is not an experiment and would be considered an observational approach.
- **Non-experimental methods are used when the study doesn't control treatment** – users' behaviour is observed without intervention. A common approach is to compare treated individuals with a control group drawn from "similar" untreated individuals. This approach, which may also be called **forensic/synthetic control**, often uses methods including matching or weighting to construct the control group.
- **Ensure test and control groups are comparable** – Groups should be equivalent across relevant characteristics, including prior behaviour and other predictive indicators, apart from the advertising treatment.
- **Clearly define the treatment and control** – The advertising treatment, target group and control condition should be clearly defined, with objective and disclosed criteria for selecting participants and controls. Groups should also be checked for contamination or unintended exposure.
- **Use reliable data and disclose potential bias** – The methodology should disclose sample sizes, assumptions, sampling and non-sampling error, potential biases and the steps taken to address them. Variables used in the experiment should be meaningful and appropriate to the use case.
- **Validate the experiment** – Validation should demonstrate that randomisation produced the intended groups, that test and control groups are comparable, and that the intended exposure/no-exposure conditions occurred. The MRC also recommends ongoing validation and documentation of the process.
- **Report uncertainty and methodology transparently** – Results should include standard error and guidance such as confidence intervals, along with disclosure of the methodology, assumptions, variables, weighting or adjustments and other relevant analytical decisions.

In essence, the MRC standard is about demonstrating that the test and control groups are genuinely comparable, the treatment is clearly defined and controlled, and the data, methodology, potential biases and results are sufficiently validated and transparent to support the conclusions drawn.

the causal-strength spectrum

Incrementality is not a single technique but a family of approaches. What unites them is the goal of isolating the true business impact of marketing activity. But not all methods are equal in their ability to establish causality, and not all use cases demand the same level of rigour. Methods that estimate incrementality vary in how well they isolate the effect of a campaign from other factors.

Methodological quality can also move a specific study up or down the evidence hierarchy. For example, a poorly implemented randomised study can be less useful than a high-quality model-based counterfactual design with excellent exposure identification, matching, sufficient sample and transparent statistical treatment.

The IAB / IAB Europe framework groups them into four families, ordered by causal strength with example approaches below (approaches are outlined in more detail in section 4):

◄ STRONGER CAUSAL EVIDENCE — WEAKER CAUSAL EVIDENCE ►

STRONG **experiment-based**	STRONG TO MODERATE **model-based counterfactual**	MODERATE TO WEAK **econometric**	WEAK **directional proxies**
Randomised Control Tests (RCTs)	Synthetic control	Market Mix Modelling (stronger causal strength when calibrated against experiments)	New-to-brand %
Hold-out / ghost ads	Natural experiment	Time-series regression	Baseline-vs-exposed
Matched markets	Machine Learning propensity models	Time series proxy with a known correlate	Platform-reported 'lift' not based on a randomised holdout
			Simple multi-touch attribution

what incrementality is not

Incrementality is distinct from a number of measurement approaches that describe activity without establishing causation. This is not to say these techniques and metrics lack value – many provide useful insight and answer legitimate ad effectiveness questions. However, the questions they answer are different from those addressed by causal experiments, so it is important to distinguish between them when interpreting results and making investment decisions.

These are some examples of where causal lift is not established:

Approach	The question it answers	What it does not establish
Attribution	Which touchpoints were involved in a conversion, and how credit is shared	Whether the advertising caused the outcome
ROAS / platform-reported return	What return was recorded against the advertising	Whether that return is incremental
Correlation	Whether outcomes moved alongside activity	Whether the activity, rather than another factor, caused the movement
Directional proxies	An indicative read (e.g. new-to-brand %, Baseline-vs-exposed, Platform-reported 'lift' metrics not based on a randomised holdout)	Causal lift, because there is no independent control to establish what would have happened without the advertising
Treatment A versus Treatment B testing	Demonstrates a difference between the two tested groups	The difference cannot reliably be attributed to the advertising itself
Incremental campaign reach	The additional unduplicated audience reached by a campaign or media channel beyond the audience already reached by other elements of the campaign or media plan	It does not mean causal incrementality in the business impact sense.
Time period comparisons	How metrics in the advertising treatment window compared to other windows.	Whether the advertising caused the difference in metrics between the periods. There are other factors that could have caused the difference.
Pre/post brand surveys	How brand metrics such as awareness, consideration or purchase intent changed between surveys run before and after the campaign.	Whether the advertising caused the change, because there is no unexposed control group to rule out other factors such as seasonality, competitor activity or wider market trends.

3 when experiments are useful: strengths, limitations and use cases

↑ back to contents page

Experiments are not the right tool for every measurement question. The table below rates how well they suit common marketing use cases, followed by a summary of their key strengths and limitations.

Experiment fit: ●●● Strong ●●○ Conditional ●○○ Limited

Use case	Experiment fit	Why experiments are (or aren't) suited
Budget allocation & planning		
Set the total level of marketing investment	●○○ Limited	MMM with response curves is the primary approach. Incrementality evidence from major channels is a supporting input used to test or calibrate the model.
Allocate budget across channels	●●○ Conditional	MMM-based scenario planning is the primary approach; channel-level experiments add causal evidence for material reallocations or uncertain channel effects.
Estimate diminishing returns / marginal ROI	●●○ Conditional	Repeated or multi-cell experiments across spend levels, or MMM response curves calibrated with causal evidence, are needed.
Allocate investment within a channel (publishers, inventory, audiences, placements)	●●● Strong	Experiments at the publisher, inventory, audience or tactic level, where feasible, show which execution options contribute the most incremental value.
Test expanded audiences, inventory or publishers for incremental reach	●●○ Conditional	Controlled tests, combined with reach, frequency and audience-overlap analysis, show whether added investment expands the effective audience or mainly increases frequency among people already reached.
Cross-channel & cross-network comparison		
Compare performance across all channels / networks	●○○ Limited	Rarely practical given ecosystem silos and fragmentation; better suited to MMM. Cross-platform ROAS comparisons are not comparable without normalising for attribution windows, identity coverage and closed-loop data.
Test whether apparent channel performance differences are causal	●●● Strong	Incrementality testing for a specific channel comparison or hypothesis— e.g. whether a channel creates incremental conversions at the level attributed reporting suggests, or whether they would have occurred anyway.

Use case	Experiment fit	Why experiments are (or aren't) suited
Size cross-channel halo effects	●●○ Conditional	Multi-cell or omnichannel geo tests can capture combined effects but require scale and careful design.
Proving value & ROI		
Validate ROI of media investment	●●● Strong	Controlled experiments are the primary approach for establishing whether an investment caused outcomes that would not otherwise have occurred. Produces incremental revenue, incremental profit, incremental acquired or retained customers. These can then be set against cost to determine iROAS or true CPA.
Demonstrate impact on sales	●●● Strong	Measures incremental conversions, customers or sales against the counterfactual.
Decide whether to increase, maintain, reduce or stop an investment	●●● Strong	Incremental return is compared with the organisation's economic threshold or decision rule. We caution that a statistically credible lift can still be too small to justify its cost.
Demonstrate impact on brand	●●● Strong	Experimentally designed brand-lift studies measure lift in awareness, consideration, preference and other attitudinal outcomes.
Testing & innovation		
Validate new channels before scaling	●●○ Conditional	Controlled experiments can establish whether a new channel creates incremental impact, but pilot budgets may be too small for stable estimates; a pilot may instead justify a larger, more credible causal test.
Test creative executions	●●● Strong	Controlled experiments under comparable audience and media conditions show which execution causes the strongest improvement in the outcome the creative is intended to influence. However, it does not necessarily show the causality of the overall media investment.
Isolate the effect of specific creative elements	●●● Strong	Controlled tests can vary the element being evaluated while holding other material elements constant. Comparing substantially different ads cannot attribute a performance difference to any one element. However, it does not necessarily show the causality of the overall media investment.

Use case	Experiment fit	Why experiments are (or aren't) suited
Test whether creative effects differ by audience, channel or context	●●● Strong	Controlled creative tests can be designed to compare results across audiences, placements, channels or contexts. However, it does not necessarily show the causality of the overall media investment.
Validate attention, completion and engagement signals	●●○ Conditional	Repeated studies or experiments can test whether changes in an intermediate creative signal correspond to changes in business outcomes; higher scores do not automatically establish greater incremental impact.
Test new tactics, targeting, frequency, formats or partnerships	●●● Strong	Controlled experiments isolate the effect of a specific change. Changing too many variables at once makes it difficult to determine what caused an improvement.
Test bid strategies, campaign types and in-flight changes	●●● Strong	Controlled experiments can show whether a change caused an improvement before it is extended. An improvement immediately following an optimisation is not proof that the change caused it.
Optimisation (in-flight & always-on)		
Optimise active and always-on campaigns	●●○ Conditional	Platform and campaign reporting provide the speed and granularity for in-flight optimisation, but you cannot always prove that your intervention drove the change unless there is a concurrent control group.
Identify which campaign components drive performance	●●○ Conditional	Controlled experiments are needed where the effect of a particular audience, placement or tactic must be isolated before drawing causal conclusions.
Calibration & validation		
Calibrate MMM and attribution	●●● Strong	Experiment results provide the causal anchor used to calibrate MMM response assumptions and attribution, and to test confidence in model estimates through triangulation with independent evidence.
Audit demand-harvesting channels (brand search, retargeting)	●●● Strong	Holdouts reveal how much demand-harvesting activity is genuinely incremental. Search often reaches consumers already expressing intent, so some attributed sales may have occurred without the advertising.

strengths

The principal strength of causal experiments is their **ability to isolate the effect of a specific marketing activity or change.** By creating a credible counterfactual—what would have happened without the intervention—experiments can provide direct evidence of whether a marketing activity caused an incremental change in an outcome.

Experiments can be used to **answer a wide range of practical marketing questions.** Depending on the design and outcome measured, they can assess incremental changes in brand awareness, consideration, purchase intent and other brand measures, as well as conversions, sales and other business outcomes. This makes experimentation useful across the marketing funnel, from understanding advertising and creative impact through to validating commercial outcomes.

A particular strength is the **ability to test specific marketing choices before scaling** them. Marketers can compare different creative executions, messages, formats, audiences, targeting strategies, media tactics, frequency levels or campaign approaches to understand which produces a stronger incremental response. Experiments can also be used to evaluate new products, features, channels or emerging advertising capabilities before committing significant investment.

This makes experimentation particularly well suited to **test-and-learn and optimisation.** Rather than simply measuring the performance of an activity after it has run, marketers can formulate a hypothesis, deliberately introduce a change, measure the incremental effect and use the result to inform the next decision. Over time, repeated experimentation can build a body of causal evidence about what works, for which audiences, under which conditions and against which outcomes.

Experiments can also provide an **independent validation point for other measurement** approaches. Results from well-designed experiments can be used to calibrate or validate MMM and attribution, helping assess whether modelled or platform-reported results are consistent with observed causal evidence.

Finally, experiments can be particularly valuable where traditional user-level measurement is constrained. Properly designed geographic or other controlled experiments can establish a counterfactual without relying on cookies or individual-level attribution, while potentially capturing outcomes that occur both online and offline. Well-designed geographic experiments should incorporate random assignment of exposed and unexposed regions and/or controls for variables such as demographics, socio-economic profile and category penetration to minimise potential bias.

> *"The core strength is causal evidence that holds up under finance and leadership scrutiny, which makes it a strong basis for investment decisions. Because geo experiments sit outside platform attribution, results are measured independently of platform algorithms and self-reported data. They also require no cookies, identifiers or user-level tracking, so the approach is privacy-safe and resilient to ongoing signal loss. Aggregating outcomes to geography brings a further benefit that user-level methods often miss: experiments can capture offline and in-store sales impact alongside online."*
>
> **— Brett Camilleri, VP of Measurement APAC at Monks**

limitations

Experiments also have significant practical and methodological constraints. They **require a sufficiently large and comparable sample, adequate spend and enough outcome volume to detect a meaningful difference.** They do require some upfront effort and time to design, while holdout groups create an opportunity cost because some audiences are deliberately withheld from the advertising. Tests may also need stable conditions around factors such as pricing, promotions, targeting and creative to avoid introducing other explanations for the observed result. Results should therefore be reported with appropriate measures of uncertainty and recognised as evidence from a particular test period and set of conditions.

Experiments are also **designed to answer specific questions rather than provide a complete view of marketing effectiveness.** A test may focus on a particular channel, tactic, audience or outcome, while businesses need to understand the combined contribution of many activities. Running multiple tests simultaneously can be difficult, particularly where channels overlap or audiences cannot be cleanly separated.

Contamination is another important risk. A control group intended to receive no treatment may still encounter the advertising through another channel or platform, reducing the ability to isolate the effect being tested. Geographic experiments can also face challenges where media cannot be cleanly controlled by location. The more channels, audiences and variables included in a test programme, the more complex and resource-intensive the design becomes.

Finally, an experiment provides evidence for the conditions and outcome tested. **Results may not automatically generalise** to different audiences, channels, creative, markets or time periods. This makes ongoing experimentation valuable but also reinforces the role of complementary approaches such as MMM and attribution in providing a broader and more continuous view of marketing effectiveness.

> *"Experiments benefit from sufficient scale and budget, so low-volume or small-budget tests can be underpowered. Geo tests need geo-targetable media and region-level outcome data; channels that cannot be cleanly geo-isolated (e.g. podcasts, linear TV, influencer activity) are better read via a time-based test. Holdouts carry some opportunity cost from foregone exposure in the control geos, mitigated by always-on designs and guardrails."*
>
> **— Peter Madani, Director of Sales and Partnerships at Lifesight**

4 choosing an approach: experiment types and how they establish incrementality

↑ back to contents page

This section describes the main experiment types and analytical methods used.

◄ STRONGER CAUSAL EVIDENCE — WEAKER CAUSAL EVIDENCE ►

STRONG **experiment-based**	STRONG TO MODERATE **model-based counterfactual**	MODERATE TO WEAK **econometric**	WEAK **directional proxies**
Randomised Control Tests (RCTs)	Synthetic control	Market Mix Modelling (stronger causal strength when calibrated against experiments)	New-to-brand %
Hold-out / ghost ads	Natural experiment	Time-series regression	Baseline-vs-exposed
Matched markets	Machine Learning propensity models	Time series proxy with a known correlate	Platform-reported 'lift' not based on a randomised holdout
			Simple multi-touch attribution

experiment-based (strong causal strength)

Randomised Control Tests (RCTs): uses randomisation to produce test and control groups. In this approach, whoever is running the RCT has the ability to explicitly control who does or does not get treatment and makes this decision truly at random. Any approach that does not control the treatment of individuals and/or does not do this at random is not an experiment and would be considered an observational approach. Useful for addressable environments with stable person-level identity and outcomes. For CTV, addressable television and household-level commercial outcomes, a household-randomised CTV test can provide strong household level evidence without establishing which person within the household watched or purchased. For bricks and mortar retail media, can also randomise stores.

Hold-out / ghost ads: a randomised experiment in which a portion of an eligible audience is assigned to a control group and withheld from the advertising treatment. Ghost-ad approaches provide a way to identify comparable control opportunities within an ad-serving environment, helping ensure that control and treatment groups are comparable.

Matched markets/Geo experiment: compares similar geographic markets, with advertising deployed in test markets and withheld from control markets, to estimate the incremental impact of the advertising on a defined outcome. Test and control markets need to be comparable on relevant characteristics, and the analysis needs to account for other factors that could affect the outcome, such as pricing, promotions,

competitive activity and other media. Particularly useful for television, radio, out-of-home, retail and integrated media activity that cannot be randomised easily at user level.

Brand lift experiment: a brand lift study, typically conducted via on-platform polls or a research panel, can use either randomised or observational exposed/control designs. Randomised treatment (where the treatment group is exposed to the advertising while the control group is withheld from the advertising, and differences in brand measures are compared to estimate incremental brand lift) provides the strongest basis for causal inference where it is feasible.

In live media environments, however, robust brand-lift measurement may use validated exposure data and carefully matched or constructed control groups. The strength of causal interpretation depends on the quality of exposure identification, group comparability and controls for potential selection bias. Where exposed/control survey designs are used, treatment and control respondents should ideally be recruited contemporaneously throughout the campaign rather than sequentially pre/post. This reduces the risk that differences in seasonality, news, competitor activity or other external influences are mistaken for advertising effects.

model-based counterfactual (strong to moderate causal strength)

Synthetic control: Creates a comparison group from untreated individuals, that are selected or weighted to closely resemble the treated group based on relevant characteristics and past behaviour. Unlike an RCT, the control group is constructed rather than randomly assigned. This means the approach can be useful where randomisation isn't possible, but there is a greater risk that unobserved differences between the groups affect the result, it approximates what would have happened to the exposed group in the absence of the intervention.

Natural Experiments: Use naturally occurring events or changes in circumstances to assess advertising impact, rather than researchers deliberately assigning treatment and control groups. For example, comparing sales in geographic markets where a campaign ran with those where it did not. They can provide evidence of incremental impact when controlled experiments aren't practical, but causal validity depends on the comparability of the groups and accounting for other influences on outcomes. Synthetic control or weighting or resampling can be used as methods to calibrate the comparison group within a natural experiment. Adding difference in difference measurement method can add further rigour.

Machine-learning propensity models: Use statistical or machine-learning techniques to estimate each individual's likelihood of being exposed to advertising or receiving a treatment, based on observed characteristics and behaviour. These propensity estimates can then be used to construct a comparable control group and estimate the outcome that would have occurred without advertising—the counterfactual. Because the control is modelled rather than randomly assigned, results can be affected by selection bias, omitted variables and data quality.

Multi-channel brand lift study: Typically involves conducting a survey via a research panel amongst a representative sample of consumers and using digital ad exposure data to identify those exposed to a

campaign across participating channels. Respondents complete a brand survey, with results compared against a matched control group who were not exposed to the campaign. Exposure classification should use the most accurate method available for the channel, ranging from deterministic/passive exposure through platform integrations to validated probabilistic or opportunity-to-see methods. The methodology and expected level of exposure certainty should be disclosed. Study design should distinguish overall campaign lift from individual-channel and combination effects where sample permits. Exposure to other campaign channels should either be controlled for explicitly or incorporated into the analytical design, rather than treated simply as unexplained contamination.

Lab forced-exposure experiment: A controlled experiment, typically conducted among a research panel, in which participants are deliberately exposed to an advertisement in a controlled research environment and their responses are compared with those of a control or comparison group that is not exposed, or is exposed to an alternative stimulus. The approach is commonly used for creative testing to assess responses such as attention, engagement, message takeout, brand perceptions, persuasion or purchase intent. Where participants are randomly assigned to treatment and control, the design can provide causal evidence of the advertising's effect on the measured response. However, because exposure takes place under controlled conditions, results may not fully reflect advertising effects in a real-world media environment.

econometric (moderate to weak causal strength)

Time-based / forecast counterfactual (incl Bayesian Structural Time Series): Estimates what would have happened without the advertising by using historical performance or trends to construct a forecasted baseline, then compares the actual outcome during the campaign with that expected baseline. The counterfactual is modelled rather than created through random assignment or a contemporaneous control group. Consequently, the result depends more heavily on the quality of the historical data, modelling assumptions and ability to account for other factors affecting the outcome.

non-causal methods (don't measure incrementality)

Pre/post brand survey or brand tracking: are generally not causal methods (do not measure incrementality). A survey, commonly conducted among a research panel, that measures brand perceptions or attitudes at different points in time, such as before and after an advertising campaign or through ongoing brand tracking. It can identify changes in metrics such as awareness, consideration, favourability or purchase intent over time. However, without a contemporaneous control group or other means of isolating advertising exposure, changes cannot reliably be attributed to advertising, as other factors may have influenced the result.

Further definitions of analytical methods or terms used in experiments are provided in the appendix.

5 designing a reliable experiment

↑ back to contents page

A credible experiment depends on good design. This section covers the key design choices — groundwork, baseline, timing, spend and measurement — and how to tailor each to the advertiser's category, purchase cycle and the outcome being measured.

laying the groundwork before the test

Before the test is built, it is worth looking at what came before it. Large-scale campaigns or tactics run in the lead-up to the testing window can carry over into the test period through delayed effects and influence the results, so what happened before the experiment is just as important as the experiment period itself. Previous campaign experience and results are also a useful input when scoping the test. Understanding what information is already available, and what has been learnt from past activity and baseline results, helps build a test that accounts for the relevant variables, such as how audiences are likely to respond.

It also pays to brief the measurement partner early and run a quick feasibility check before the media plan is set. This should confirm practical requirements such as geo availability, minimum campaign duration and spend, the outcome data required, exposure and holdout capability, and any other campaign activity that will need to be controlled. It gives planners an early read on what can be measured while the plan is still being built, rather than discovering measurement constraints once activity is already booked.

The publisher or platform being tested is another valuable source of insight. Publishers can share what they know about their audience, such as whether people spend more time with them than with other publishers, which can guide how the test is built and provide a lens for interpreting the results. Finally, the creative should link clearly to the business outcome being measured. If the experiment is measuring response to a call to action, that call to action needs to appear in the creative.

establishing a reliable baseline

The baseline provides the benchmark for estimating what would have happened without the advertising. If it does not reflect normal business performance, the measured lift may be misleading.

For sales-based experiments, the pre-period should be long enough to capture normal trading patterns, typically around four to 12 weeks depending on the category. Unusual events, including promotions, price changes, competitor activity and supply disruptions, should be accounted for.

For matched-market experiments, the test and control markets should also demonstrate similar sales patterns before the intervention begins. This is known as a *parallel-trends check*. If their performance was already diverging, differences observed during the test may not be caused by the advertising.

Where possible, use a full 52 weeks of historical sales data to assess market comparability, capturing seasonal and promotional patterns across the year. If less data is available, use the longest comparable period and disclose the limitation.

For synthetic or time-based control approaches, the baseline is modelled from historical data rather than observed in a live control group. The quality and relevance of that data are therefore particularly important.

setting the intervention period

The intervention period is the time during which the test group receives the advertising and the control group does not. The test must run long enough to detect an effect that matters to the business, while remaining short enough to minimise the influence of other changes in the market.

Before launch, the experiment should establish its *minimum detectable effect (MDE)*: the smallest lift it is designed to reliably detect. This depends on the variability of the outcome, the number of users or markets included, and the required statistical confidence. Larger samples, longer tests and more consistent outcomes can help detect smaller effects. The MDE should be calculated before the test begins, so the design can be assessed against the business objective.

As a guide, geo experiments typically require at least four to six weeks of continuous advertising to accumulate sufficient exposure and outcome data. The actual duration should be determined by the test's statistical requirements and the stability of the environment.

Throughout the test, material conditions such as pricing, promotions, creative and media weight should remain consistent between test and control markets, wherever possible. Changes in these conditions can introduce alternative explanations for the results.

The test should also be designed to deliver results in time to inform the business decision it is intended to support.

setting the right level of media spend

Media spend needs to be sufficient to generate a measurable effect, but not so high that the test operates at an unrealistic level of investment.

Too little spend may produce an effect that is too small to detect. Too much may push the channel into saturation, where additional spend generates diminishing returns. The resulting lift may then be difficult to apply to lower or higher levels of investment.

Spend should be determined alongside the test duration, sample size, variability of the outcome and expected effect. These factors are interdependent and may need to be adjusted together to achieve a credible and commercially relevant test.

The level of spend tested should also reflect the investment decision being considered. Results from a high-spend test should not automatically be assumed to apply at lower levels of investment.

measuring the post-intervention effect

Advertising can continue to influence behaviour after a campaign ends. This carryover effect, or *halo*, may contribute to the total incremental impact of the campaign.

Whether to include a post-intervention measurement period should depend on the objective of the experiment and the advertiser's purchase cycle.

For frequently purchased products, advertising effects may dissipate within days or weeks. For longer consideration categories, brand-led campaigns, subscriptions or repeat-purchase activity, effects may persist for weeks or months.

If the objective is to measure the full incremental return of the campaign, the post-period should be included and its length based on the expected duration of the carryover effect. If the objective is to measure the impact during the campaign only, a post-period may not be necessary.

Where a post-period is included, it should be long enough to detect a meaningful effect. A short or underpowered post-period may fail to identify carryover, even when it exists.

> *"I say this to our agency partners all the time; a bad result is better than no result. A test is only as credible as the design that underpins it. Get the baseline wrong and you measure noise; cut the intervention short and you measure nothing; ignore the halo and you measure only part of the story. In my experience, the tests that change decisions are not the ones with the biggest lifts. They are the ones where the design is clean enough so the lift can actually be trusted."*
>
> **— Ben Oliver, Media & ICP Lead ANZ, Mars Petcare**

considerations across purchase cycles

The time between advertising exposure and purchase can vary substantially by category. Effects may become apparent within days for frequently purchased products but may take months or longer for categories such as automotive or financial services. While infrequent purchase categories can present different opportunities for detecting incremental effects, this depends on the ability to reach a sufficient number of in-market consumers and generate adequate sample sizes. Category purchase cycles have important implications for experiment design, measurement windows and the interpretation of results.

	Shorter purchase cycles	Longer purchase cycles
Examples	Regularly purchased goods (e.g. milk, shampoo); FMCG; direct-response journeys	Infrequent purchases (e.g. TVs, sofas); automotive; financial services
Measurement window	Shorter observation and outcome windows may be sufficient to capture incremental purchases.	Longer observation and outcome windows may be needed to capture delayed purchases and advertising carryover.
Statistical power	Higher outcome volumes can support experiments with shorter observation periods	Fewer, higher-value outcomes may require larger samples and longer observation periods
Main risk	Habitual or pre-existing demand being attributed to advertising	Long lags create more opportunity for competing influences during the decision period

	Shorter purchase cycles	Longer purchase cycles
Stronger evidence	Large-scale randomised or geo experiments supported by purchase history	Longer-horizon experiments, complemented by MMM and brand evidence

- **shorter purchase cycles**
  - **Separate existing demand from new demand:** While you might observe strong lift among existing buyers, this lift may be short-lived while the lift among new-to-brand has sustained impact. Where possible, assess results separately for existing and new buyers.
  - **Account for promotions and pricing:** These can be important sources of competing influence, particularly in FMCG where discounting is common. This is why experiments that isolate the pure effect of advertising is critical.
  - **Set and justify the measurement window upfront:** Experiment duration should be appropriate to the category's purchase cycle, supported by evidence where possible, and disclosed before the campaign runs.
  - **Allow for repeat purchasing:** Advertising may influence purchase frequency or retention, not just generate a single additional purchase.
- **longer purchase cycles**
  - **Treat the measurement window as an assumption:** There is no universally correct window. Longer windows may capture more of the advertising effect but also introduce more competing influences.
  - **Plan for statistical power:** Low observed conversion or an inconclusive result should not be interpreted as evidence that advertising had no effect. It is important to plan an experiment that has the requisite treatment intensity.
  - **Measure intermediate outcomes:** Brand lift, site visits or adviser contact can provide evidence of progress before a purchase occurs, although these should not automatically be treated as equivalent to sales outcomes.
  - **Disclose extrapolation:** If a short-period result is used to estimate a longer-term effect, the basis for the extrapolation should be disclosed and periodically reassessed.
  - **Consider brand maturity:** Established brands with high awareness may be harder to shift on core brand measures than newer or less established brands.

6 considerations for measuring incrementality by channel

↑ back to contents page

The practical challenges of designing and interpreting experiments vary by channel, depending on how advertising is delivered, how exposure is identified, the ability to establish a comparable control group and the availability of outcome data. Other factors, such as existing consumer intent, audience overlap, exposure outside the test environment and competing marketing activity, can also affect the reliability and interpretation of results.

Exposure classification should use the most accurate method available for the channel, ranging from deterministic or passive exposure through platform integrations to validated probabilistic or opportunity-to-see methods. The methodology and expected level of exposure certainty should be disclosed.

ILLUSTRATIVE EXAMPLE

Ways to measure ad exposure by channel for brand lift

> *"Kantar looks at ad exposure for brand lift studies, not as one universal construct, but specific to the media channel. Distinction is crucial because misclassification of exposure directly affects the validity of a brand-lift result."*
>
> **— Andreea Chirila, Client Development Director, Media ROI at Kantar**

Kantar brand lift studies use validated exposure data and carefully matched or constructed control groups. The strength of causal interpretation depends on the quality of exposure identification, group comparability and controls for potential selection bias.

The ways Kantar define exposure for brand lift studies differs across channels:

- Digital: passive/tag-based exposure can be used.
- Direct platform integrations: passive exposure information for supported environments can be used.
- Social/walled gardens: probabilistic approaches such as Probable Media Exposure can be required.
- TV: passive and/or opportunity-to-see approaches can apply.
- OOH, radio and print: exposure can rely on OTS measures combined with media consumption and campaign delivery information.

For multi-channel campaigns, study design should distinguish overall campaign lift from individual-channel and combination effects where sample permits. Exposure to other campaign channels should either be controlled for explicitly or incorporated into the analytical design, rather than treated simply as unexplained contamination.

The following considerations highlight channel-specific practices and watchouts for designing, conducting and interpreting experiments. They complement, rather than replace, the core principles of experimentation discussed in this paper.

- **social media**
  - **Algorithmic delivery and optimisation:** Platforms use automated algorithms to determine audience selection, ad delivery and frequency. This makes it important to understand how exposure and control groups are established, particularly where delivery is optimised towards users most likely to engage or convert.
  - **Organic versus paid exposure:** Users may encounter the same creator, brand or content organically as well as through paid advertising. Experiments need to distinguish the incremental contribution of paid exposure from existing organic engagement and familiarity.
  - **Engagement-driven exposure:** Social platforms encourage sharing, commenting and resharing, meaning content can extend beyond the original paid audience. This can create exposure within control groups and make it harder to maintain a clean exposed-versus-control distinction.
  - **Platform-specific exposure definitions:** Platforms define and report impressions, views, video views and engagement differently. Understand what constitutes an advertising exposure within the experiment, particularly for short-form video, autoplay and interactive formats.
- **paid search**
  - **High-intent audiences:** Paid search often reaches consumers who are already expressing intent, so some attributed sales may have occurred without the advertising. Holdout experiments help reveal how much of this demand-harvesting activity is genuinely incremental.
  - **Test an optimised campaign, not a set-up phase:** An incrementality test run on an unoptimised campaign, such as one using manual bidding or a fragmented keyword structure, is likely to measure temporary set-up inefficiencies rather than the channel's actual potential. This can understate the incremental value of paid search and lead to the wrong investment decision.
  - **Establish a stable baseline before the test begins:** In line with the baseline principles in section 5, the campaign should be running in a stable, fully optimised state before the experiment starts. In paid search, this typically means using automated bidding strategies such as target cost-per-acquisition (tCPA) or target return on ad spend (tROAS).
  - **Separate the advertising effect from bidding noise:** An unstable bidding system adds variation that can be mistaken for, or can mask, the effect of the advertising. A stable, optimised starting point means the experiment measures true incremental value rather than this background noise.
- **CTV/BVOD/streaming**
  - **Household versus individual exposure:** CTV can be viewed on shared household screens, so household-level assignment may be more appropriate than individual-level assignment in some studies. The correct unit depends on the campaign and available identity data.
  - **Representation of audience that can be tracked by individual platforms or their measurement partners:** understanding whether the people that can be tracked and can be included in the experiment (the 'panel') are any different to those who can't be tracked as part of the experiment.

- **Cross-platform fragmentation:** Audiences are spread across broadcasters, streaming services, apps, devices and platforms, each with different data access and reporting capabilities. People in a control group may encounter the same campaign through another streaming service, linear TV or online video, reducing the measured difference between groups.

ILLUSTRATIVE EXAMPLE

Randomised ghost-ad holdouts for CTV conversion and brand outcomes

Ashley Spinks, Senior Measurement Partner Australia at Netflix Ads, outlines how Netflix uses randomised ghost-ad holdouts to measure the incremental conversions and brand lift driven by advertising on its platform.

Netflix's first-party lift studies give advertisers a clear question every campaign faces: what impact did our advertising spend have that wouldn't have happened anyway? A randomly assigned group is held back from seeing a particular ad, an audience who otherwise would have been exposed to the campaign. These members form the "control" group, and we record the ad they would have seen, known as a "ghost ad". Netflix Ads viewers who were exposed to the ad fall into the "test group". Both groups pass through the same eligibility and auction, so the two groups are comparable by design.

Conversion Lift is used when an advertiser needs to understand did ad exposure drive more actions? We compare how many people in the test group took an action such as purchases, sign-ups and leads, with how many in the control group did. Events arrive through first-party server-to-server and tag-based pathways using hashed identifiers, deduplication and consented data only. Each study begins with a feasibility check on impressions, duration, baseline conversion rate and minimum detectable effect, targeting 80% power. It then runs with a post-flight observation window for lagged conversions and reports one lift for the full window, with a confidence interval and p-value.

Brand Lift unveils findings like did the ad change how people think about the brand? After an episode or movie ends, an eligible Netflix Ads viewer may be asked a short 1-question survey. Each person answers just one question and together the answers pool to form a broader brand lift survey. It compares exposed members who saw the ad with held-out control members drawn from the same pool of viewers. The difference between the two groups assesses the impact the ad had on recall, awareness, consideration and intent.

By comparing viewers who saw an ad with a randomly held-back group who didn't, Netflix shows advertisers what their campaigns genuinely drove in actions and brand perception shifts.

ILLUSTRATIVE EXAMPLE

Measuring entertainment outcomes (tune-in)

Yasmin Sanders, VP ANZ & SEA at Samba, explains how Samba measures whether advertising for a show or movie gets more households to watch it.

When a publisher promotes a new show or movie, the outcome that matters is whether people watch it. Samba calls these entertainment outcomes, also known as "tune-in". A publisher's randomised control group is the ideal comparison, but it often doesn't survive the move to a third-party measurement panel. Some control households can't be matched, and others turn out to have seen the ad anyway through another service or screen. The control group no longer reflects the original design.

Samba's approach, CaML (Causal Machine Learning), gives each exposed household its own control. Using viewing history, demographics, location and past engagement from the Samba household panel, a model predicts how likely that household would have been to watch the content without seeing the ad. Lift is the gap between what households actually did and what the model predicted.

Because CaML can be layered on top of a publisher's holdout, it sits across both "experiment-based" and "model-based counterfactual" in the framework in section 4. It is a good example of real-world measurement blending methods. It is best suited to entertainment outcomes for streaming and linear TV premieres.

- **retail media**
  - **High-intent audiences and selection bias:** Retail media often targets shoppers who are already browsing, searching or considering a purchase. They may be more likely to buy regardless of advertising, which is why it is even more critical to design experiments with control groups.
  - **Price, promotions and product availability:** Sales can change because of discounts, stockouts, distribution, product placement or other retail activity, independently of the advertising being tested, which is why it is even more critical to design experiments with control groups.
  - **Overlapping retail touchpoints:** Onsite ads, offsite media, sponsored search, retailer apps, CRM and in-store activity can overlap, making it difficult to isolate the effect of individual tactics.
  - **Substitution and halo effects:** Advertising may shift purchases between products, brands or categories, rather than generate additional sales overall. Experiments need to consider both the advertised product and relevant wider outcomes.
  - **Retailer-specific measurement:** Retailers can link advertising exposure to transaction data, but results may be limited to one retailer and may not capture purchases elsewhere. Attributed sales alone do not establish incremental impact.
  - **Different retailer methods and data:** Retail media networks vary in their data, definitions, reporting and experimental capabilities, making results difficult to compare across retailers and networks.

- **Brought-forward sales:** It is important to measure the post-treatment window of time to see if the marketing just brought forward sales or, more hopefully, sustained sales lift even after the treatment was withdrawn.

ILLUSTRATIVE EXAMPLE

Measuring the cross-channel sales impact of retail media

Ali Khan, FMCG Vertical Head at Amazon Ads, outlines an example of an experiment using a geographic holdout measuring in-store sales lift for a Nestle campaign.

For the 2025 launch of Nescafé Espresso Concentrate, Nestlé wanted to know whether Amazon Ads drove extra supermarket sales, not just sales on Amazon.

What ran - Prime Video shoppable ads, display, Amazon Influencers and a Brand Store, aimed at 18- to 34-year-olds. The campaign reached 3.2 million unique viewers.

How it was measured - One state was kept dark as a control. Stores in campaign regions were matched with similar control stores, and an independent provider compared their in-store sales during and after the campaign.

What it found - The featured range grew 8.9% nationally above the control during the campaign. Brand measures also moved: consideration rose 20.7%, purchase intent 8.7% and recall 10.3%.

Why it matters - Closed-loop retail media doesn't have to stop at the retailer's own sales. A simple geographic holdout can show whether retail media drives sales across the wider grocery market. It also shows where it doesn't: here, for example the media had little effect during half-price promotion weeks.

Limitations - The control market was chosen, not randomly assigned, so this is a matched-market design rather than a randomised experiment.

- **digital OOH**
  - **Incomplete exposure records:** not all exposed individuals are identifiable in DOOH exposure records, models must account for potential overlap between test and control groups, particularly when using lookalike modelling or control groups based on observed behaviours. Unintended inclusion of exposed individuals in control groups can reduce accuracy in determining true lift.
  - **Modelled exposure:** DOOH has no device-level impression, so exposure is usually inferred by matching mobile location data to a screen's viewable area and play times. This indicates opportunity to see rather than confirmed viewing.
  - **Location-driven selection bias:** Screens near stores or in high-traffic areas reach people who may already be likely to visit or buy. Simple comparisons of exposed and unexposed audiences can overstate the effect.

- **Control contamination:** Screens sit in shared public spaces and people often pass several of them, so control audiences or markets can be exposed without being detected. Control screens or markets should be matched on audience, traffic and proximity to the outcome location.
- **Delivery and external factors:** Programmatic pacing, dayparting and pricing mean actual plays can differ from the plan. Weather, events, seasonality and other media in the same areas can also affect outcomes, independently of the advertising being tested.
- **Panel scale and data linkage:** Measurement often relies on mobile location panels and probabilistic matching to web, app or sales data. Declining device signals, smaller panels in some markets and low baseline outcomes can limit statistical power. Attributed outcomes alone do not establish incremental impact.
- **Different measurement approaches:** Media owners and measurement providers vary in exposure methods, control construction, match rates and reporting. Geo or matched-market tests can reduce reliance on device data but need comparable markets and longer durations, making results difficult to compare across providers.
- **Two primary methodologies are recommended for incrementality measurement:** Synthetic Control and Matched Market / Store testing.

- **streaming audio**
  - **Logged-in environments support true randomisation:** Serving ads to authenticated listeners provides greater control over experimentation, enabling user-level holdout and ghost-ad designs.
  - **Delayed effects and measurement windows:** Audio advertising can have carryover effects that influence behaviour beyond the campaign period. Measurement windows, including a post-campaign read, need to be long enough to capture these effects, while balancing the increased noise and contamination risk of longer tests.
  - **Solo vs. co-listening**: Experiments should consider whether exposure and outcomes are measured at the individual or account/household level, especially for campaigns delivered across both loudspeaker (smart speakers, in-car) and headphone devices. Shared-account use raises a similar consideration, where the person exposed may not be the person whose outcomes are measured.
  - **Don't mistake audience reach as measurable reach:** Whether it be through exposure or opportunity to hear, finding robust sample can be a challenge when other audience variables are overlayed. Work with the publisher on what sample sizes might be achievable.
  - **There is a delayed impact for action with audio:** Many people listen to audio whilst they are driving or at work and often don't have capacity in that moment to go to a website or purchase.
  - **Understand the campaign that sits around audio:** Audio often forms part of a broader multimedia campaign. Consider the control group and where they might be exposed to marketing through other channels.
  - **If running a geo experiment, consider the audio market:** For radio, in some markets there are more stations than others. This impacts audience share, R&F and investment. Don't assume the media schedule for one market immediately works in another.

- **Consider the impact of premium environments:** Audio's effectiveness is going to vary between what assets are utilised. For example, live reads are usually more effective than brand spots. When building the experiment, be clear on the assets being tested and overall outcome of the study.

- **podcasts**
  - **Don't mistake audience reach as measurable reach:** Whether it be through exposure or opportunity to hear, finding robust sample can be a challenge when other audience variables are overlayed. Work with the publisher on what sample sizes might be achievable. Some titles are not big enough to find sample of in research panels.
  - **Podcasts are intimate so profile of the control group needs to be considered:** Podcast audiences may differ from non-listeners in interests, attitudes and behaviours that are also predictive of brand outcomes. Ensure your control group are also podcast listeners to titles of a similar genre to what you are testing.
  - **Podcast campaigns are more effective when the brand and title are a natural or contextual fit:** Podcast effectiveness can vary substantially by title, host, audience and the natural fit between the brand and content. Results should be interpreted at the level tested rather than automatically generalised to podcast advertising overall. This might help narrow down which titles you might want to use as part of the experiment.
  - **Testing the title or testing the platform:** A host read ad compared to a standard produced ad are different experiences and deliver differences in effectiveness. Understanding whether you are testing the title, the host or the platform broadly helps determine the most appropriate way to run the experiment.

- **creator / influencer**
  - **Isolate the incremental impact of paid amplification:** Where creator content receives paid support, boosting can be evaluated much like other forms of paid media. Comparing creators or content with similar baseline engagement and follower profiles, but different levels of paid support, can help isolate the additional impact generated by boosting, rather than attributing all performance to the creator. To establish incrementality, the comparison also needs to account for differences in audience, content and campaign conditions. Ideally, paid support would be tested through controlled exposure and holdout groups.
  - **Brand-lift surveys can be executed to provide critical upper-funnel validation:** exposed-versus-control survey studies used to measure awareness, recall, and consideration.
  - **Measurement today is directional but not definitive:** the ecosystem does not yet consistently support incrementality-based creator evaluation. The opportunity ahead is not simply to measure creators better, it is to integrate creators fully into the financial and planning systems that govern modern media investment. Engagement is only an input signal, and earned media value (EMV) doesn't measure sales or incrementality.
  - **Attribution under-credits creator influence:** last-click links, promo codes and affiliate tracking reward what closes the sale. Creators mostly work earlier, in discovery, consideration and trust.

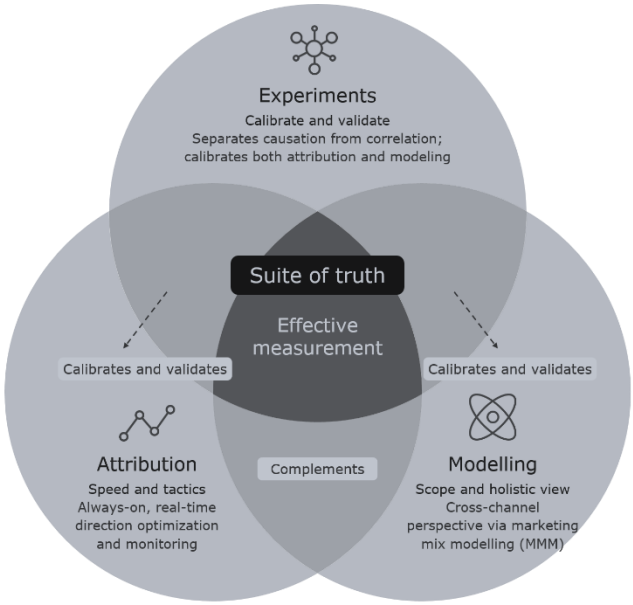
- **Data quality and verification:** delivery data is mostly self-reported by platforms, auditing is limited, and fake followers and bots persist.
- **Link creator exposure to sales where data permits:** first-party, CRM and retail media data can connect exposure to sales where data-sharing agreements allow.
- **Move towards full-funnel, integrated measurement:** build creators into cross-media measurement alongside MMM, this is the most mature stage of assessing impact of creator activity.

7 putting experiments in context: how they work in the measurement toolkit

↑ back to contents page

Experiments are most valuable as the causal anchor of a broader measurement system, and used in combination with MMM, attribution and brand measurement. Each methodology has value and a role to play when employed strategically as part of a wider framework that balances rigour with practicality.

MMM provides a strategic, cross-channel, long-horizon view, and attribution provides granular, near-real-time optimisation signals, but neither establishes causation on its own. Experiments do, and their results can be fed back into both to calibrate MMM priors, saturation and return curves, and to de-bias attribution.



Source; Meta (IAB MeasureUp 2026 presentation 'Proving ROI in an Evolving Marketing Ecosystem')

> *"For experiments to play this calibration role reliably, a few conditions matter. The hypothesis and success criteria should be set before launch and tied to a real business decision. The counterfactual must be credible, for example validated matched markets in a geo test, and the design should have adequate statistical power confirmed up front. The test environment needs to stay stable (consistent pricing, promotions and creative), and KPI definitions should be aligned across experiments, MMM and attribution so results are actually comparable. Testing also works best at a regular cadence, so learnings continually inform the models rather than sitting in one-off studies."*
>
> **— Brett Camilleri, VP of Measurement APAC at Monks**

> *"Lifesight views robustly conducted experiments as the calibration engine and source of causal truth. Experiments establish ground-truth lift; that lift can then calibrate MMM saturation, ROI curves and adjust attribution multipliers, so all three layers share one causal foundation. The methods are complementary: MMM provides always-on, long-horizon, all-channel decomposition and forecasting (including hard-to-isolate channels); attribution provides granular, real-time optimisation; and experiments validate and recalibrate both, resolving disagreements between them."*
>
> **— Peter Madani, Director of Sales and Partnerships at Lifesight**

why results across MMM, attribution and experiments may not align

Different methods often produce different answers. Kantar research cited in Meta's *Building a Suite of Truth* found 55% of advertisers have experienced contradicting results across solutions.

Differences between methods do not necessarily indicate that one method is wrong. Before comparing results, establish whether the methods are measuring the same intervention, population, outcome and time period, and whether they are answering the same causal question. Where they differ, the reasons for the difference should be understood rather than simply averaging the results or assuming that agreement proves accuracy.

Common reasons your results may diverge across methods include:

- **They answer different questions:** MMM, attribution and experiments measure different aspects of advertising effectiveness and may have different objectives or demands.
- **They use different counterfactuals or none at all:** An experiment compares outcomes with and without an intervention, while attribution often describes observed customer journeys and MMM infers relationships from historical variation.
- **They measure different populations:** An experiment may cover a specific campaign, audience, platform or market, while MMM may represent the broader business and attribution may only cover identifiable users or journeys.
- **Exposure may be defined differently:** An "exposed" consumer can receive different creative, formats, frequency, placements or durations. Methods may also use different rules for determining whether someone was exposed.
- **Identity and data coverage differ:** Methods may have different levels of visibility into audiences, exposures, conversions and cross-channel activity. Identity matching can also exclude or misclassify some activity.
- **Targeting and prior intent can influence results:** People who are more likely to purchase may also be more likely to receive or engage with advertising. This can make observed attribution differ from incremental impact.
- **Time periods and measurement windows differ:** Results can change depending on the campaign period, seasonality, observation window and when outcomes are measured.
- **Short-term and longer-term effects are treated differently:** Experiments may capture effects within a defined test window, while MMM can account for carryover and longer-term effects. Attribution is often focused on observable actions closer to conversion.
- **Other factors are treated differently:** MMM may explicitly account for factors such as price, promotions, distribution, seasonality and economic conditions, whereas other methods may have less visibility of these influences.
- **Model assumptions and methodology differ:** MMM results depend on model specification, transformations, lag structures, priors and available variation. Attribution depends on its attribution rules and algorithms, while experiments depend on their design and implementation.

- **Interventions and creative may not be comparable:** A campaign-level experiment, a channel-level MMM estimate and a user-level attribution result may relate to different combinations of media, audience, creative and tactics. Comparing them directly can therefore be misleading.
- **Uncertainty and variation are inherent:** Sampling variation, statistical uncertainty, model uncertainty, imperfect implementation and genuine differences in response across audiences and situations mean that estimates will not necessarily converge.

using brand lift experiments and sales lift experiments in tandem

Brand lift and sales lift experiments are discrete methodologies that answer different questions. **Sales lift** assesses whether advertising generated incremental sales or conversions, while **brand lift** assesses whether advertising changed intermediate outcomes such as awareness, consideration or purchase intent.

There may, however, be value in running them in tandem when there is a clear rationale for understanding both **whether advertising generated a commercial outcome and whether it changed brand measures that may precede or contribute to that outcome**. This can be particularly relevant for campaigns with longer purchase cycles, where brand effects may emerge before sales can be observed. Running both does not mean that the results should be expected to align or that a brand effect can be assumed to explain a sales effect.

If running them in tandem, consider:

- **Use the same campaign and audience:** Where possible, run both studies against the same campaign, target audience and geographic market so differences are less likely to reflect differences in population or media activity.
- **Align the exposure period:** Use the same campaign dates and, where possible, the same exposure definition and treatment period.
- **Use comparable treatment and control groups:** Apply consistent principles for constructing treatment and control groups, including eligibility, randomisation and exclusion criteria.
- **Align exposure measurement:** Use the same definition of advertising exposure where possible, recognising that the data requirements for measuring exposure may differ between brand and sales outcomes.
- **Align the measurement period with the expected effect:** Set measurement windows that reflect the expected timing of each outcome. They may not be identical, but any differences should be understood and documented.
- **Use the same campaign conditions:** Avoid introducing material differences in creative, media mix, targeting, frequency or campaign timing between the populations being assessed.
- **Control for relevant external influences:** Where possible, account for factors such as promotions, pricing, seasonality or other activity that could affect either brand measures or sales.
- **Pre-specify the measures and hypotheses:** Define the primary brand and sales outcomes, expected direction of effect and analysis approach before the campaign begins.

- **Apply appropriate statistical power to each outcome:** Size each study to detect the expected effect for its particular outcome. A sample that is sufficient for a brand measure may not be sufficient to detect incremental sales.
- **Document differences between the studies:** Where methodology, exposure measurement, sample, timing or outcome definitions cannot be aligned, document those differences so they can be considered when interpreting the results.

using experiments to calibrate attribution and MMM

Calibration refers to the process of adjusting and refining measurement models, algorithms, or systems to ensure they accurately reflect real-world outcomes. This involves comparing predicted results with actual results, identifying discrepancies, and making adjustments to improve the accuracy and reliability of marketing metrics.

Source: Meta, "Building a Suite of Truth: A Hybrid Approach to Measuring Incrementality"

calibrating attribution

Calibration adjusts your attribution so it reflects each channel's true, incremental impact — measured against a Conversion Lift result.

Meta's 2026 MeasureUp presentation outlined an approach to calibrating attribution by creating a multiplier, a relatively quick way to ensure accuracy, particularly if you use multiple marketing channels and have the ability to run incrementality experiments on each.

Let's say you run ads on Meta and other channels. Using an MTA model, you determined the conversions driven by your channels and Meta drove 100. By running a more rigorous Conversion Lift test on Meta, you see that Meta actually drove 150 conversions, a 50% increase over what the attribution model showed. To align your MTA model with incrementality, you would calculate a multiplier for Meta using the following calculation:

When you run an incrementality experiment to determine a channel's true incremental conversions, you can divide that number by the conversions shown in your MTA model to arrive at a multiplier.

Apply the multiplier to your MTA results to align them with incrementality. Refresh the multiplier periodically for accuracy.

ILLUSTRATIVE EXAMPLE

$$\frac{\text{Incremental conversions}}{\text{Attributed conversions}} = \frac{150}{100} = 1.5\times \text{ multiplier}$$

Source: Meta, IAB MeasureUp 2026 presentation "Proving ROI in an Evolving Marketing Ecosystem"

This multiplier is then applied to Meta's results from the attribution model going forward. In this case, to adjust Meta's results they would use the following: 100 × 1.5 = 150 conversions.

However, if the total conversions after applying the multiplier exceed *actual* total conversions across all channels, it's best to repeat this process for other platforms. Run platform experiments for each channel and use that same calculation to get multipliers for each.

The multiplier can also be a number less than 1. Using the same method outlined above, attributed sales can be 100, while the experiment says that only 80 sales are truly incremental, 20 of those sales would have happened anyway and therefore the multiplier is 0.8.

Multipliers are just one method for calibrating attribution models, and they should be updated periodically to stay accurate.

> *"Adopting incrementality doesn't require an all-or-nothing shift. Attribution still plays an important role—it's fast, always-on, and useful for directional optimisation and monitoring performance in near real time. But it shouldn't be treated as the single source of truth, because it can't reliably separate correlation from causation.*
>
> *The strongest strategy is to use incrementality to calibrate and validate what attribution (best for fast, day-to-day optimisation decisions) and Market Mix Modelling (best for cross-channel, longer-horizon measurement) are telling you—creating a more complete view. We call this approach a "suite of truth": a dynamic suite of trusted tools and methods that can provide a more reliable framework for evaluating media effectiveness."*
>
> **— Eby Jose, Marketing Science Partner at Meta**

calibrating market mix modelling

Market Mix Models provide a cross-channel understanding of business growth by estimating the sales generated by each media channel. However, if used alone, MMM can misattribute results, systematically over- or under-estimating the effectiveness of specific channels. For that reason, it's important to calibrate MMM.

Meta's *Building a Suite of Truth* report says "There are a few approaches that advertisers can take to validate or calibrate their MMM with experiments, depending on their objectives and data availability. Calibrating 'qualitatively' by confirming that experiments and MMM results are directionally similar, is relatively straightforward which can help increase the confidence in model outcomes. More advanced and sophisticated approaches include using studies to choose between different candidate models or incorporating experiment results into the model by setting or constraining channel coefficients using incremental estimates."

ILLUSTRATIVE EXAMPLE

Meta's 2026 MeasureUp presentation outlined three approaches, from basic to advanced. Watch the presentation.

Compare results from MMM and experiments	Guide MMM optimisation using experiments	Fine-tune model details with experiments
If both methods show similar trends, the MMM is considered reliable. This is easy to do, but not very precise.	Set the average impact from experiments as a constraint in MMM. This improves accuracy without the need to rebuild the model.	Comprehensive parameter calibration uses experiment results to adjust model details like ad stock and media saturation predictions.

◄ EASIER TO IMPLEMENT LESS RIGOROUS

HARDER TO IMPLEMENT ► MORE RIGOROUS

Source: Meta, IAB MeasureUp 2026 presentation "Proving ROI in an Evolving Marketing Ecosystem"

calibrating market mix modelling with geo experiments

Because Market Mix Modelling (MMM) relies on historical observational data, it is inherently susceptible to multicollinearity and selection bias—particularly for always-on channels like Paid Search that scale in tandem with baseline demand. Integrating independent, randomised geographic experiments into the model structure anchors observational parameters to causal reality.

Geographic experiments act as a reality check for broader marketing models. Modern open-source modelling frameworks allow advertisers to feed the results of a geographic test directly into the model to guide its calculations. The model treats the experiment as an established piece of truth, adjusting its curves to align with the real-world lift measured during the test.

The strength of this guidance depends entirely on the quality of the experiment. A highly precise test with a clear, statistically strong result will tightly guide the model, preventing it from overestimating a channel's value. Conversely, a test with high uncertainty or volatile data will offer looser guidance, giving the model more flexibility to rely on historical trends.

ILLUSTRATIVE EXAMPLE

Historically, translating the results of a geographic test into an input that a marketing model could understand was highly complex and prone to manual error. Recent open-source tools, such as Google's geographic experimentation library Meridian GeoX, have standardised this process. These libraries help advertisers design matched-market tests, isolate the true incremental sales driven by a campaign, and output those results in a standardised format that can be easily plugged into modern modelling tools. This makes calibration an accessible, repeatable process rather than a complex, manual task.

calibrating MMM in practice

Guidance from Prophet on how calibration works in a Bayesian MMM

Many modern MMMs, including widely used open-source frameworks, are built on Bayesian methods. This makes calibration a natural fit. An experiment result becomes prior information about a channel's effect, and the model combines it with the evidence in the historical data.

The experiment's uncertainty matters as much as its point estimate. A tightly estimated, well-powered lift result should constrain the model strongly. A wide or borderline result should only nudge it. Encoding the experiment's confidence interval in the prior keeps the model honest about what is and isn't known. Fixing a coefficient to the point estimate does not.

A more direct approach treats the experiment as an additional observation the model must explain. Rather than setting a prior on a channel's overall return, the model is required to reproduce the lift observed when spend changed during the test period. This anchors the response curve at the spend level actually tested.

> *"Calibration works best as a conversation between methods rather than an override. An experiment tells you very precisely what happened in one place, at one spend level, over a few weeks. The model tells you how that fits into the whole business over years. Encoding the experiment with its uncertainty, checking whether the model could have predicted it, and then using the model to choose the next test is what turns one-off studies into knowledge that compounds."*
>
> **— Rachel Scott, Head of Decision Science, Prophet**

translating an experiment into model terms

An experiment and an MMM rarely measure exactly the same thing, so a result needs translating before it is used. Five things need aligning:

- **The outcome.** A platform conversion lift result and an MMM built on total sales across online and offline channels are not the same quantity.
- **The time horizon.** Experiments read effects within a fixed window, while MMM accounts for carryover. A short test will typically understate the total effect of channels with long-lasting impact, unless the post-period is extended or the unobserved tail is accounted for.
- **The spend level.** An experiment measures response at the spend level tested. It locates one point or segment on the response curve, not the whole curve. Extrapolating well above or below that level still relies on the model's saturation assumptions. Multi-cell tests at different spend levels are the most valuable for informing diminishing returns.
- **The scope.** If a test ran in selected geographies or audiences, the assumptions used to scale it to the national business should be disclosed.
- **The conditions.** Seasonality, creative, pricing and promotions during the test may differ from the period the model is estimating.

validation, not just calibration

Experiments are also a test of the model. Before using a result as an input, compare it with what the uncalibrated model predicts for the same intervention. Agreement within the bounds of uncertainty builds confidence. Disagreement is diagnostic. It can point to correlated media plans the model cannot separate, a misspecified carryover structure, or a missing variable. Holding some experiments back from calibration and using them as out-of-sample checks gives an honest read of how accurate the model actually is.

keeping calibration current

Experiment results are evidence from a particular time and set of conditions. Weight results by how recent, well-powered and relevant they are, and allow older results to carry less influence over time. Re-test after material changes in creative, strategy, platform algorithms or spend level, rather than letting a single historical test define a channel indefinitely.

watchouts

- An inconclusive experiment is not evidence of zero effect. Calibrating a channel to zero on the basis of an underpowered test will understate its contribution.
- Any study, whether designed, run and analysed by a platform, publisher, vendor or in-house team, should be scrutinised before they are used to calibrate. Completion of the IAB Australia Transparency Template for Incrementality Experiments (section 9) helps here.
- Calibrating every channel tightly leaves the model little room to learn from the data. It also means an error in any one experiment flows straight through to budget recommendations.
- Model outputs should make clear which channel estimates are anchored to experiments, so users understand where the evidence comes from.

closing the loop: using MMM to design the next experiment

Calibration should run in both directions. A Bayesian MMM's uncertainty shows where knowledge is weakest relative to the spend at stake. Channels with large budgets and wide uncertainty are where the next experiment will be most valuable.

The model can also support test design. It can estimate expected effect sizes for power calculations, help select comparable markets, and account for baseline variation. The result is a continuous cycle: model, identify uncertainty, test, calibrate, reallocate. Each experiment improves the model, and the model makes each experiment more useful.

8 note on changes to privacy legislation

↑ back to contents page

On 31 August 2026, the Australian Government released an exposure draft of the Privacy Amendment (Personal Data Protection) Bill 2026 and an accompanying consultation paper. The draft Bill is the second tranche of reforms to the Privacy Act 1988 (Cth) and the most significant overhaul of Australian privacy law in more than a decade. **Please note the legislation is in draft form so further changes are likely.**

The proposed reforms could have implications for how marketers collect, share and use data for advertising measurement. The proposed broader definition of personal information, strengthened expectations around meaningful consent, and a new 'fair and reasonable' test for handling personal information will place greater emphasis on understanding data provenance, the purposes for which data was originally collected, and whether its subsequent use for measurement is appropriate and transparent. Data sharing between advertisers, platforms and measurement providers will also require careful consideration, particularly where personal information is made accessible to third parties. **The proposed law currently requires consent whenever personal information is 'traded' (i.e. disclosed to a third party)**. While this does not mean consent will be required for every measurement activity, marketers should urgently audit what data they hold, where it comes from, what permissions and consent apply, and whether their measurement practices can be conducted without using personal information, without disclosure to a third party, or using completely de-identified data.

IAB Australia has prepared a summary (available to IAB members) setting out each proposed change and what it means in practice for members. These are proposed reforms, not yet finalised legislation.

The key areas of reform include:

- a new 'fair and reasonable' test on how personal information is used applying alongside consent requirements
- targeted advertising is explicitly being brought into the direct marketing framework, including audience, segment and cohort targeting based on personal information and online behavioural data.
- strengthened opt-out rules triggered on targeted advertising, plus simpler, clearer notifications
- strengthened consent measures to ensure consent is voluntary, informed, current, specific and unambiguous when collecting personal information and for when using it
- the definition of personal information is expanding from information "about" an individual to information that "relates to" an identified or reasonably identifiable person
- a new principle of the right to erasure on large digital platforms.

IAB Australia will be providing further guidance when the legislation is finalised.

9 transparency template for incrementality experiments

↑ back to contents page

This template gives measurement practitioners a standardised way to understand how advertising effectiveness experiments are set up and executed to interpret results with confidence. The template draws together the causal requirements and disclosure standards referenced throughout this paper. It is intended to travel with a study's results so that the two are read together.

The template can also be adapted as a briefing form at the campaign and measurement planning stage. Completing the purpose, scope and data fields with planned budget, flighting and markets lets measurement partners assess feasibility while the media plan is still flexible, rather than discovering constraints once activity is booked.

Field	Details (to complete)
Purpose & scope	
Study name	
Advertiser / brand *The business the study is run for.*	
Provider / analyst *Vendor or in-house team responsible for design and analysis*	
Dates *Dates of pre-period, campaign and read-out.*	
Desired business outcome *What the advertiser/brand is ultimately trying to accomplish.*	
Business decision *The specific decision that measurement needs to support in pursuit of that outcome.*	
Measurement question *What must be learned in order to make the business decision*	
Channel(s) / tactic(s) tested *The campaign, media platforms, channels, tactic, creative, audiences and markets to be tested.*	

Field	Details (to complete)
Metrics & data	
Outcome metric definition *Definition of the outcome metric and how it is measured.*	
Data sources *Where the data is coming from, period of historical data use and data granularity.*	
Privacy & consent *How data collection, linkage and sharing met applicable privacy, consent and data-use requirements.*	
Methodology	
Experiment type *RCTs / Conversion Lift, hold-out / ghost ads, matched markets, synthetic control, ML propensity model, or time-based.*	
Counterfactual *How the what-if-not scenario was created.*	
Test / control design *Number of cells, treatment/control split and holdout size.*	
Comparability basis *How test and control were matched or balanced, including prior purchase history.*	
Analysis method & assumptions *Model used to estimate lift and the key assumptions the causal claim relies on, e.g. parallel trends or no spillover.*	
Statistical robustness	
Sample base & size *Sample sizes in control and test groups.*	
Confidence & significance *Confidence level, confidence interval and significance of the result.*	

Field	Details (to complete)
Error *Sampling and non-sampling error; standard error.*	
QA checks *Contamination, exposure-validation and data-quality checks completed.*	
Known limitations & caveats	
Limitations, risks, biases *Platforms and channels covered and not covered.* *Identified risks and the mitigations applied.* *Point-in-time validity and recommended refresh date.* *Residual biases and the controls applied.* *Uses the result does not support.*	
Actions & recommendations	
Investment implications & retesting *What the result means for investment, e.g. scale, maintain, reallocate or reduce spend.* *How confidently the result can be applied to future planning, and where further evidence is needed.* *Whether and when the experiment should be retested, e.g. at a set refresh date or after changes to creative, media mix, pricing or market conditions.*	

ILLUSTRATIVE EXAMPLE

Applying the IAB Australia Transparency Template for Incrementality Experiments in practice – entertainment advertiser case study by Foxcatcher

Foxcatcher's incrementality case study examines whether digital advertising campaigns for a ticketed live entertainment advertiser generated incremental ticket sales. It compares households exposed to the campaign with statistically matched, unexposed households to estimate the impact on ticket purchases and revenue.

Not only does the case study outline the results but it also provides the necessary transparency on methodology, sensitivity checks for robustness and the limitations that will allow the client to fully understand the context of the results.

Access the full case study here.



A appendix — more definitions and terms

↑ back to contents page

Definitions of analytical methods or terms used in experiments

Difference-in-differences: is an analytical method (not an experimental design in itself), it compares the change in an outcome over time between a test group exposed to a marketing activity and a comparable control group that was not exposed. The difference between those changes is used to estimate the incremental effect of the activity. It can be applied to a matched-market test or other observational setup.

Propensity Score Matching (PSM): A statistical method used to construct a comparable control group when advertising exposure has not been randomly assigned. Each individual's likelihood of being exposed to advertising is estimated using observed characteristics and prior behaviour, and exposed individuals are matched with similar unexposed individuals. Differences in subsequent outcomes between the matched groups can then be used to estimate incremental impact. Because matching can only account for observed factors, unmeasured differences between groups may still introduce bias.

Regression Discontinuity (RD): A quasi-experimental method used to estimate incremental impact when advertising treatment is assigned according to a defined threshold or cut-off. Outcomes for people or other units just above and below the threshold are compared, on the assumption that they are otherwise sufficiently similar. A discontinuity in the outcome at the threshold can provide evidence of a causal effect of the advertising. The approach is most applicable where a clear treatment threshold exists and typically estimates the effect for units close to that threshold.

CUPED (Controlled-experiment Using Pre-Experiment Data): A variance-reduction technique applied to controlled experiments that uses relevant pre-experiment data, such as historical purchases or conversions, to account for predictable variation in the outcome being measured. By reducing unexplained variance, CUPED can improve the statistical precision of the estimated treatment effect and increase the ability to detect incremental impact. CUPED does not establish causality itself; it is used to improve the efficiency and precision of a properly designed experiment.

Stepped-wedge design: An alternative experimental design that can be used where a phased rollout is appropriate. Groups, such as geographic markets or audience segments, are introduced to an advertising intervention at different points in time, with all groups eventually receiving the intervention. Outcomes are measured before and after each group receives the advertising, while groups yet to receive the intervention provide a contemporaneous comparison. Where the order of introduction is randomly assigned, the design can provide causal evidence of incremental impact while allowing the advertising to be rolled out across the full target population.

Matched pairs design: Commonly used in store-by-store comparisons. For every test store, there is control store and each pair is compared. The lift between test and control is the average or median of pairwise comparisons.

B appendix — find out more

↑ back to contents page

Want to find out more? The resources below are a good starting point for anyone looking to build their understanding of incrementality and advertising effectiveness measurement. They include industry frameworks and standards from IAB US, IAB Europe, the Media Rating Council (MRC), and other industry bodies.

more from the IAB Advertising Effectiveness Council

- IAB Australia, The essential marketing effectiveness reading list (Sep 2026)
- IAB Australia, Future of Measurement Report 2026 (Sep 2026)
- IAB Australia, Market Mix Modelling Landscape Report (2025)
- IAB Australia, Marketing Measurement Innovation Series, Module 5: Ad Effectiveness Through Continuous Experiments (2024)

further reading

- Media Rating Council, Outcomes and Data Quality Standards (Sep 2022).
- IAB US, Cross-Channel Measurement Guide (Oct 2026).
- IAB Australia, Australian Retail Media Measurement Principles and Guidance (2024).
- IAB US & IAB Europe, Guidelines for Incremental Measurement in Commerce Media (Nov 2025).
- IAB Europe, Connected TV Measurement Framework and Transparency Principles (2026).
- IAB US, The As-Is Measurement Landscape in the Creator Economy (Jan 2026).
- IAB US, Digital Out-Of-Home (DOOH) Measurement Guide (2025).
- CIMM, Measuring What Advertising Changes: Attribution, Incrementality and the Future of Outcomes Measurement (Sep 2026).
- IAB US, State of Data 2026: The AI-Powered Measurement Transformation (Feb 2026).
- IAB US, Modernizing MMM: Best Practices for Marketers (Dec 2025).
- Meta, Building a Suite of Truth: A Hybrid Approach to Measuring Incrementality (May 2025).

iab. australia