

cimm

Coalition for Innovative
Media Measurement

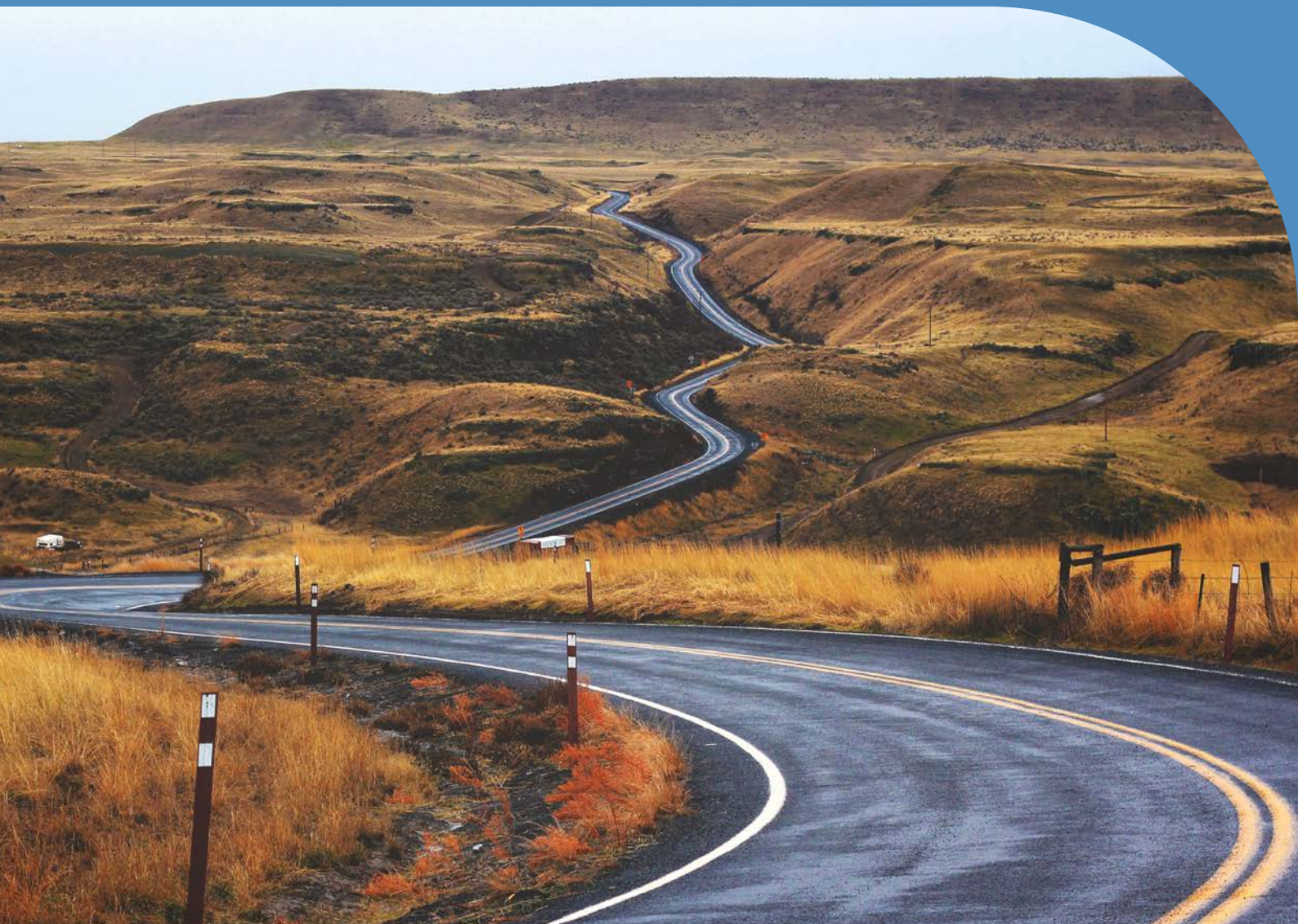


A CIMM Working Paper

Measuring What Advertising Changes:

Attribution, Incrementality and the
Future of Outcomes Measurement

September 2026





About CIMA

The Coalition for Innovative Media Measurement (CIMA) is a non-partisan, pan-industry organization focused on improving media measurement, currency, data collaboration and the use of emerging metrics. CIMA brings together advertisers, agencies, media owners, platforms, technology companies and measurement providers to support research, testing, dialogue and the development of practical industry approaches to shared measurement challenges.

This paper was authored by Jon Watts, Managing Director of CIMA.

Research Objectives and Approach

CIMM Working Papers examine emerging measurement issues, synthesize evidence and identify questions and potential areas for further industry work.

They are discussion documents and do not necessarily represent final CIMM positions. These papers are intended to provide context, develop perspectives, catalyze debate and identify questions for further consideration during our program. They do not necessarily represent a final CIMM position; rather, they are designed to inform discussion and help lay the groundwork for future CIMM research and industry workstreams.

CIMM identified attribution and outcomes measurement as one of its three principal strategic priorities for 2026. Television and video have already developed substantial capabilities to connect advertising with consumer and business outcomes, and those capabilities are becoming increasingly important to measurement, optimization and commercial decision-making. This Working Paper considers what is required for the next stage of that development. It examines two connected challenges for the U.S. television and video marketplace: how to strengthen the evidentiary basis of attribution and outcomes claims, and how to make credible outcome signals easier to interpret and use across a fragmented, multi-seller marketplace. It reviews what current systems measure, what their outputs can support, how outcomes are becoming embedded in media execution and commercial practice, and where targeted cross-industry work could improve transparency, experimental practice and marketplace interoperability.

The objective is to provide a practical roadmap for strengthening the marketplace's existing capabilities, not to prescribe a single methodology, provider or commercial model.

Scope, Method and Evidentiary Status

This paper is an analytical and industry-policy Working Paper examining attribution, incrementality and outcomes measurement in the U.S. television and video marketplace. It is not an accreditation assessment of any individual measurement service, a legal opinion, or an empirical estimate of market-wide advertising effectiveness.

The analysis draws on published academic and industry research, measurement standards and guidance, publicly available information about current market practices and capabilities, CIMM research and industry activity, and consultation and review with practitioners and subject-matter experts. Its primary focus is U.S. television and video, while drawing selectively on evidence from digital advertising, experimentation, marketing-mix modeling and other adjacent fields where they illuminate the underlying measurement questions.

The paper distinguishes between observed facts, published evidence, analytical interpretation and proposals for future industry work. The priorities identified in Chapter 11 are areas for consideration and further investigation; they do not represent agreed industry commitments or imply endorsement by organizations or individuals consulted during the development of the paper. Throughout, the strength and granularity of any conclusion should be understood as bounded by the underlying data, methodology, assumptions and research design.



Acknowledgements

This paper represents the independent analysis and synthesis of the author. It has benefited substantially from discussions with CIMA members and other participants across the advertising, television, measurement, data and technology ecosystem, as well as from detailed review and feedback from experts in advertising effectiveness, attribution, experimentation and causal measurement.

The author is grateful to everyone who contributed perspectives, challenged assumptions, reviewed drafts and helped strengthen the analysis. Participation in these discussions or reviews should not be interpreted as endorsement of every conclusion or proposal contained in the paper.

Special thanks are due to Rick Bruner, CEO of Central Control, for his feedback, ideas and support.

All errors, judgments and interpretations remain the responsibility of the author.

A note on this paper

This is deliberately a substantial Working Paper, intended both as an industry analysis and as a detailed reference resource. Readers seeking a shorter route through the material can use the one-page Key Takeaways, the industry-roadmap exhibit and the Executive Summary that follow.

CIMA will also publish a series of shorter papers in the coming months, drawing out selected themes, findings and practical implications from the research.

Key Takeaways

Television and video have developed sophisticated capabilities for connecting advertising with consumer and business outcomes. The next stage is not simply to generate more outcomes data, but to strengthen what that evidence can establish and make credible outcome signals easier to interpret and use across the marketplace.

1 Connection is not causation. Linking an advertising exposure with a subsequent search, visit, purchase or other outcome establishes an observed relationship. Attribution can assign credit for that outcome. Establishing incrementality requires a credible estimate of what would otherwise have happened.

2 More data and greater granularity do not automatically produce stronger causal evidence. Better identity resolution, richer consumer-path data and more detailed reporting can improve observability and operational usefulness without eliminating selection effects, prior purchase propensity, missing influences or other sources of bias.

3 The observed consumer journey is not necessarily the causal journey. Consumers bring existing preferences, intentions and purchasing histories to advertising, while price, availability, recommendations, competitor activity, previous brand experience and creative execution can all affect outcomes. Even a rich sequence of touchpoints may therefore provide only part of the explanation.

4 No single methodology provides the complete answer. Attribution, randomized and quasi-experiments, MMM, brand research and customer evidence illuminate different dimensions of effectiveness. They should inform and test one another, while evidence with stronger causal identification should carry greater weight for causal claims.

5 AI can strengthen the measurement system, but it cannot eliminate the identification problem. AI can integrate evidence, interrogate assumptions, identify inconsistencies and accelerate learning. It cannot manufacture the missing counterfactual or turn association into causation through greater computation alone.

6 TV's next opportunity is to combine stronger evidence with greater marketplace usability. The ecosystem already possesses extensive outcomes capabilities. Greater consistency in definitions, disclosures, experimental practice, data interfaces and cross-seller workflows could make those capabilities easier to use at scale without requiring a single methodology, provider or commercial model. A federated marketplace offers a more practical direction than a universal outcomes system or guarantee.

An industry roadmap:

Three priorities for practical cross-industry work

The paper identifies three connected opportunities to strengthen television and video outcomes measurement: improving the clarity of outcomes claims, strengthening a particularly important form of causal experimentation, and reducing unnecessary friction in using credible outcome signals across sellers.

An industry roadmap for TV and video outcomes measurement

Priority	What it is intended to achieve	Practical output
1. Establish common definitions and disclosure requirements for outcomes measurement	Make material outcomes claims easier to understand and compare by specifying what was measured, what was inferred, the data and population involved, the methodology and counterfactual supporting any causal claim, and the relevant assumptions, uncertainty and granularity.	A common outcomes-measurement disclosure framework , including an Incrementality Claim Specification , designed for use in briefs, RFPs, contracts, methodology documentation and reporting.
2. Develop and validate common practice guidelines for geographic randomized experiments in television and video	Establish clearer expectations for commissioning, designing, executing, analyzing and interpreting geo-RCTs, while building directly on existing MRC, ARF, IAB, academic and practitioner guidance.	A practical Geo-RCT Practice Framework , supported by common commissioning, design, power-analysis, treatment-integrity and reporting templates and tested against real TV and video experiments.
3. Assess how to make the television outcomes marketplace more consistent, scalable and interoperable	Identify where common or compatible definitions, interfaces, disclosures and workflows would materially reduce friction when advertisers use outcomes capabilities across multiple sellers, without imposing methodological or commercial uniformity.	A stage-gated feasibility assessment and multi-seller testing program identifying where interoperability creates demonstrable value and whether new standards, shared interfaces or other industry coordination are warranted.

Contents

Executive Summary	8
Guide to Terminology	15
Introduction	17
Chapter 1: Foundational principles: what attribution can – and cannot – tell us	18
Chapter 2: What exactly are we measuring and what are we inferring?	34
Chapter 3: How consumer journeys shape the attribution challenge	46
Chapter 4: Putting it all together – the measurement confidence gap	53
Chapter 5: The U.S. marketplace today – abundant signals, uneven proof	58
Chapter 6: Addressable TV and CTV: a measurement revolution – or a fragmentation problem?	66
Chapter 7: From attribution to incrementality – building credible causal evidence	73
Chapter 8: Beyond the next transaction – measuring brand and long-term effects	86
Chapter 9: From measurement systems to measurement intelligence	96
Chapter 10: From outcomes capability to marketplace scale	109
Chapter 11: Priorities for practical cross-industry work on TV attribution and outcomes measurement	125
Conclusion: From connection to confidence	153
Appendix: Sources and References	154

Executive Summary

Television and video have developed a sophisticated ecosystem for measuring what happens after advertising is delivered. Exposure can now be connected with an extraordinary range of consumer and commercial outcomes, and those signals increasingly influence planning, optimization and commercial arrangements. The next challenge is to establish more consistently what those connections mean – including when they support a claim about what advertising actually caused.

Exposure can now be connected with searches, visits, leads, subscriptions, purchases, retention and other forms of business value at a scale that would have been difficult to imagine a decade ago. Household-, device- and account-level data can be linked with advertiser first-party data, digital behavior and retailer transactions. Addressable television, connected television and streaming provide richer exposure records and new opportunities for experimentation. Data clean rooms allow datasets controlled by different organizations to contribute to common analyses. Outcome signals increasingly feed directly into targeting, bidding, optimization and, in parts of the market, commercial guarantees.

Together, these developments have expanded the range and operational use of television and video outcomes measurement.

As outcomes evidence becomes more consequential to budget allocation, optimization, guarantees and media buying, two questions become more important: whether the evidence can establish what advertising changed, and whether credible outcome signals can be used consistently across a fragmented, multi-seller marketplace.

That tension is already visible in the market. In iSpot's 2026 survey of agency and advertiser executives, more than 45% identified business outcomes as the most critical factor in buying and negotiating media, even as 47% described measuring those outcomes across linear and streaming television as a significant challenge¹.

Major integrated platforms such as Amazon and Google can combine media delivery, identity, first-party signals, conversion data, measurement and optimization within closely connected environments. This can reduce data fragmentation and shorten the path from measurement to action, but it does not automatically strengthen causal identification or provide a complete view of behavior and outcomes beyond the platform. The strength of any measurement claim still depends on its data, measurement boundary, comparison strategy, research design and counterfactual.

The wider TV and video ecosystem has developed through a more distributed structure, involving multiple publishers, platforms, identity systems, data providers and measurement approaches. Its opportunity is to improve consistency and interoperability across those systems while preserving their different strengths, commercial models and methodological approaches. Existing industry initiatives, including the MRC's 2022 Outcomes and Data Quality Standards, provide important methodological foundations for this work, although implementation, disclosure and adoption remain uneven.

But the ability to connect advertising exposure with a subsequent outcome does not necessarily establish how advertising contributed to that outcome.

The paper starts from a central proposition: The observed path to transaction is not necessarily the consumer's complete journey, and the complete journey does not necessarily reveal the full causal process.

Consumers bring existing preferences, previous purchasing, category needs and varying degrees of purchase intent to the advertising they receive. Targeting can concentrate advertising among consumers already likely to respond.

¹ Source: iSpot, *The Short List: Key Takeaways from iSpot's 2026 Video Ad Strategy Survey* (May 2026).



Media consumption, demographics, geography and market conditions can influence both advertising exposure and purchasing. Important advertising and non-advertising influences may remain outside the observable data. Advertising effects can also accumulate across exposures, interact across channels and develop over periods extending well beyond the immediate transaction.

Advertising exposure is also not a homogeneous treatment. Different creative executions can produce materially different levels of attention, emotional response, memory, brand association and behavioral response. Where creative varies across audiences, publishers, platforms or campaign conditions, observed differences in outcomes may reflect differences in what was shown as well as differences in who was reached or where the advertising appeared. Credible outcomes measurement needs to identify material variation in creative alongside variation in audience, media and exposure.

Connection establishes what was observed. Attribution assigns credit for an observed outcome. Incrementality asks what changed because the advertising occurred. These are different claims, and they require different evidence. Better identity resolution can strengthen confidence that exposure and outcome records belong together; it does not, by itself, establish causality. More granular attribution can improve operational insight without necessarily producing equally granular causal knowledge.

Two connected questions follow. First, what do the available data and methodologies allow the market to conclude about what advertising changed, with what degree of confidence and at what level of granularity? Second, what common definitions, disclosures, experimental practices and interfaces could make credible outcome signals easier to interpret and use across the fragmented U.S. television and video marketplace?



The core measurement problem

The central measurement problem remains the counterfactual: what would have happened if the advertising had not occurred?

Exposure is rarely random. Consumers who see advertising may differ from those who do not in prior purchasing, brand preference, category need, media behavior or predicted likelihood to buy, while targeting can concentrate advertising among consumers already expected to respond. Richer consumer-path data can improve understanding of the observed sequence without necessarily distinguishing pre-existing demand from advertising response or capturing all of the influences that produced the outcome.

Different methodologies address different parts of this problem. Attribution can provide granular evidence about observable pathways; randomized experiments can provide stronger causal evidence about defined interventions; quasi-experiments can extend causal inference where randomization is unavailable; MMM can provide broader cross-channel and longer-term context; and brand, customer and retailer evidence can capture effects outside immediate transactions. The objective is not one definitive attribution number, but a connected evidence system in which the scope and strength of each conclusion remain visible.

Where the ecosystem stands today

Against this framework, the contemporary U.S. attribution and outcomes ecosystem is technically sophisticated but uneven in the evidentiary strength of the claims it can support.

The market is becoming far more capable of connecting television and video delivery with identities and commercial outcomes; analyzing advertiser, publisher and retailer data together; conducting experiments in some platform, addressable, streaming and geographic environments; and incorporating outcome signals directly into optimization. Modern MMM, Bayesian approaches and experiment-informed attribution create further opportunities to connect previously separate forms of evidence.

The resulting challenge is no longer simply one of data availability. It is one of evidentiary confidence. The resulting measurement-confidence gap arises when the industry's capacity to generate detailed outcomes data exceeds its capacity to support equally detailed causal conclusions.

A system may report performance by publisher, audience, household, placement or impression even when credible causal evidence exists only at the campaign, channel or market level. Greater granularity can be operationally valuable, but it should not be mistaken for greater causal confidence.

The common language of "outcomes" can further obscure these distinctions. A website visit is an outcome. A purchase is an outcome. So are attributed and incremental purchases – but they do not represent the same kind of evidence. "Outcome" describes what is being measured; it does not, by itself, establish what can be inferred about advertising's contribution to it.

The industry is highly capable of generating and connecting outcome signals, but less consistent in defining what those signals establish, distinguishing attribution from incrementality, communicating uncertainty and matching reported granularity to the strength of the supporting evidence.

The principal challenge is not simply to create more measurement. It is to make the measurement already available more transparent, interpretable and appropriately matched to the decisions it is intended to inform.

Where outcomes measurement is heading

The analysis points to two related challenges for the next stage of outcomes measurement.

The first is **evidentiary**: how to combine attribution, experiments, MMM, brand research and customer evidence without claiming greater causal confidence or granularity than those methods support.

The second is **operational**: how to make credible outcome signals easier to activate, measure, optimize and compare across a television marketplace distributed among multiple sellers, identity systems, measurement providers and data environments.

A connected evidence system combines methods according to the questions and levels of inference they can support. Randomized experiments should provide the strongest anchors for causal claims where they can be validly implemented; quasi-experiments can extend causal inference where randomization is unavailable; observational attribution can provide speed, scale and granularity; MMM can provide broader cross-channel, business and longer-term context; and brand and customer evidence can capture effects that are not visible in immediate transactions. Their outputs can inform, test and constrain one another without being treated as causally equivalent.

This implies a practical hierarchy of evidence for causal claims. Where validly implemented, randomized evidence should generally carry the greatest weight in determining whether advertising caused an incremental change. Credible quasi-experimental designs can extend causal inference where randomization is unavailable, subject to stronger identification assumptions. Observational attribution and predictive models can provide valuable evidence about pathways, associations, heterogeneity and likely response, but they should not acquire causal status simply because they are granular, sophisticated or accurate at prediction. Methodological pluralism should not imply methodological equivalence.

AI can accelerate this learning system by helping to integrate evidence, interrogate assumptions, identify inconsistencies and support model criticism and hypothesis generation. It does not change the underlying requirements of causal inference: more data or computation cannot create an unobserved counterfactual or turn association into causation. The objective is a coherent portfolio of evidence in which the scope, assumptions and uncertainty of different estimates remain visible and stronger causal evidence is given greater weight for causal claims.

On the operational side, television already supports substantial outcomes measurement, optimization and, in parts of the market, outcome guarantees. The unresolved question is how efficiently those capabilities operate when advertisers buy across multiple sellers.

An advertiser may currently need to reconnect first-party data, reconcile different outcome and exposure definitions, interpret different identity and attribution approaches, normalize reporting and establish separate mechanisms for using outcome signals in optimization. Individual systems can function effectively while the collective workflow remains costly and fragmented.

This creates two related requirements:

1. **Infrastructure interoperability** concerns whether advertiser data, exposure records, outcome signals and reports can move efficiently across sellers.
2. **Measurement interoperability** concerns whether the resulting outputs are sufficiently transparent and comparable for advertisers to understand what each result means. Technical connectivity does not automatically create measurement equivalence.

The objective is to provide sufficient consistency in selected definitions, interfaces, disclosures and workflows to make credible outcomes capabilities easier to use across television while preserving legitimate competition and differentiation.

That distinction also places an important limit on what marketplace scale should be expected to mean. A universal television outcomes guarantee – under which an advertiser could buy one guaranteed business result seamlessly across the full linear, streaming and CTV ecosystem – would require far more than common data connectivity. Participating sellers would need sufficiently compatible definitions of the outcome and exposure, identity and matching arrangements, measurement populations, attribution rules, evidentiary standards, reporting, contractual responsibility and remedies. Where an incremental outcome was guaranteed, the parties would also need a credible basis for establishing the counterfactual. Those requirements make a single pan-industry guarantee considerably more demanding than an interoperable outcomes marketplace.

A more practical model is federated. Participating sellers could use common or compatible infrastructure to receive advertiser outcome signals, connect those signals with exposure data and produce measurement that is easier for advertisers to interpret and use across the market. Individual companies would still retain their own inventory, pricing, identity approaches, technology, measurement partners and commercial models. Where it makes commercial sense, groups of sellers could also develop joint or comparable outcome-based offerings without requiring the whole market to adopt one model.

Such an approach could support a package-level guarantee without implying that the separate causal contribution of every publisher or impression within the package has been identified.

A practical industry roadmap

The analysis points to three opportunities for practical cross-industry work: establishing common definitions and disclosure requirements for outcomes measurement; developing and validating common practice guidelines for geographic randomized experiments in television and video; and assessing the minimum interoperability required to make credible outcome signals easier to use across multiple television sellers.

These priorities deliberately address different layers of the problem. The first improves the meaning and interpretability of outcomes claims; the second strengthens practice around an important source of causal evidence; and the third addresses the operational friction created by a fragmented marketplace. None requires a single measurement methodology, provider, identity system or commercial model. Together, they offer a practical roadmap for strengthening and connecting capabilities that television and video have already built.

The opportunity is to support the development of a more mature, integrated marketplace, one that provides stronger evidence about what advertising really changed, combined with greater consistency in how credible outcome signals can be understood and used across the marketplace. The proposals are summarized in the industry-roadmap exhibit at the front of this paper and developed in detail in Chapter 11.

Guide to Terminology

Various terms used in attribution and outcomes measurement require careful distinction. The definitions below indicate how they are used throughout this Working Paper.

Term	Meaning in this paper
Outcome	A measured result following advertising activity. Outcomes can include intermediate responses such as search or website visitation; brand effects such as awareness or consideration; behavioral outcomes such as purchases or subscriptions; and economic outcomes such as revenue, margin or customer value. An outcome does not, by itself, establish that advertising caused it.
Attribution	The process of assigning credit for an observed outcome to one or more advertising or marketing touchpoints under a defined rule or model. Attribution can be useful without necessarily establishing incrementality or causation.
Incrementality	The estimated change in an outcome caused by a defined advertising intervention relative to a specified alternative or counterfactual. An incrementality claim should make clear what changed, relative to what, for whom, against which outcome, within what surrounding media environment and over what period .
Counterfactual	An estimate of what would have happened in the absence of the advertising intervention being evaluated, or under a specified alternative intervention. The difference between the observed or estimated outcome under treatment and the counterfactual is the basis of an incremental causal estimate.
Exposure	A recorded opportunity for advertising to have been delivered or encountered. Exposure can be measured at different units – including device, household, account or person – and the unit should be stated explicitly. Recorded exposure does not necessarily establish that an individual actually attended to the advertising.
Creative execution	The advertising content delivered within an exposure, including its message, branding, imagery, sound, narrative, offer, format and other executional characteristics. Creative can influence whether an exposure generates attention, emotion, memory, persuasion or behavioral response. Differences in creative execution can contribute materially to differences in observed outcomes and should be distinguished, where relevant, from differences in audience, media delivery and context.
Identity resolution	The process of determining which devices, accounts, households or other records are believed to belong together. Better identity resolution can improve the accuracy of data linkage, but deterministic identity should not be confused with deterministic causality .
Experiment	A study in which treatment is assigned according to a defined experimental design in order to create a credible comparison between advertising conditions. Where assignment is randomized, the paper distinguishes what was randomized – for example advertising eligibility – from what was subsequently observed, such as actual exposure.
Treatment effect	The estimated difference in outcomes caused by a defined intervention relative to its counterfactual. The precise estimand depends on the design. For example, an experiment that randomizes advertising eligibility may estimate an intent-to-treat effect rather than the causal effect of an individual impression.
Causal evidence / causal estimate	Evidence or an estimate intended to establish how an outcome changed because of an intervention rather than merely whether exposure and outcome were associated. The strength of a causal claim depends on the research design, counterfactual and assumptions supporting identification.

Term	Meaning in this paper
Identification strategy	The feature of a research design or analytical framework used to distinguish the effect of advertising from other factors that could have produced the outcome. Examples include random assignment, credible quasi-experimental variation or explicitly specified counterfactual assumptions. Predictive accuracy, data granularity or model complexity are not, by themselves, identification strategies for causality.
Granularity	The level of detail at which a result is reported – for example campaign, channel, publisher, audience, household or impression. Greater granularity does not necessarily imply greater causal confidence. The paper distinguishes the resolution of the output from the resolution at which the evidence can credibly support the claim .
Marketing mix modeling (MMM)	Statistical modeling that uses aggregate variation in media activity, business outcomes and other demand drivers to estimate the contribution of marketing across channels and over time. MMM can provide broad cross-media and longer-term context, but remains dependent on data quality, available variation, model specification and causal assumptions.
Outcomes-based optimization	The use of observed, attributed, predicted or causal outcome signals to influence subsequent media delivery – for example audience selection, bidding, frequency, inventory or budget allocation. Optimization against an outcome does not, by itself, establish that the optimized outcome is incremental.
Outcome guarantee	A commercial commitment in which a seller agrees to deliver a specified level of performance against a defined outcome metric. The guarantee should identify the outcome, measurement methodology, eligible population, measurement window and whether the result is observed, attributed or incremental. An outcome guarantee does not necessarily imply outcomes-based pricing.
Multi-seller outcome guarantee	A commercial commitment in which a defined group of participating sellers collectively supports a specified outcome commitment across a combined media package. The guarantee may be measured at the level of the package without establishing the separate causal contribution of every participating seller or impression. Any such arrangement should specify the participating inventory, outcome definition, eligible population, measurement methodology, identity and data arrangements, treatment of overlap, evidentiary status of the metric, allocation of commercial responsibility and remedy if the commitment is not achieved.
Outcomes marketplace	The set of commercial and measurement arrangements through which business outcomes inform television and video planning, buying, measurement, optimization, valuation or guarantees. In this paper, a more interoperable outcomes marketplace does not imply a single outcomes currency or common methodology; it refers to sufficient consistency in definitions, data interfaces, disclosures and workflows to allow outcomes capabilities to operate more effectively across sellers.

Throughout the paper, these distinctions support a common principle: **the language, scope and granularity of a measurement claim should not exceed what the underlying evidence can reasonably establish.**

Introduction

Television measurement has moved well beyond delivery alone. The industry has developed substantial capabilities for connecting advertising exposures with subsequent consumer and business outcomes across linear television, streaming and CTV. Outcome signals increasingly inform planning, optimization, measurement and commercial arrangements as well as retrospective reporting.

That progress creates a new set of questions. The ability to observe and connect more behavior does not necessarily produce equivalent confidence about what advertising caused. At the same time, the U.S. television and video market remains distributed across multiple sellers, identity systems, data environments and measurement approaches, making credible outcome signals more difficult to use consistently across the marketplace than within any one system.

This paper therefore addresses **two connected challenges**. The first is **evidentiary**: what can different forms of outcomes evidence reasonably establish about what advertising changed, with what confidence and at what level of granularity? The second is **operational**: what forms of consistency and interoperability could make credible outcome signals easier to interpret, activate, measure and use across multiple television and video sellers?

The paper starts from an important distinction. Observing that an outcome followed an advertising exposure is different from establishing that the advertising caused the outcome. Attribution, incrementality, experimentation, MMM, brand research and customer evidence answer different questions and support different forms of inference. The purpose of the paper is not to identify a single “best” methodology, but to clarify those differences, examine how the methods can operate as a connected evidence system and consider how the resulting evidence can support a more mature outcomes marketplace.

The chapters that follow move from foundational measurement principles and consumer journeys through the contemporary U.S. marketplace, addressable television and CTV, causal experimentation, long-term and brand effects, AI-enabled measurement and the development of television as an outcomes platform. The final chapter translates the analysis into three priorities for practical cross-industry work.

About this Working Paper: scope and status

CIMM Working Papers examine emerging measurement issues, synthesize available evidence and identify questions and potential areas for further industry work. They are discussion documents intended to inform debate and future CIMM activity rather than statements of an agreed industry position.

This paper is an analytical and industry-policy assessment of attribution, incrementality and outcomes measurement in the U.S. television and video marketplace. It is not an accreditation assessment of an individual measurement service, a legal opinion or an empirical estimate of market-wide advertising effectiveness. The analysis draws on published academic and industry research, measurement standards and guidance, publicly available information about current practices and capabilities, CIMM research and industry activity, and consultation with practitioners and subject-matter experts. Its primary focus is U.S. television and video, drawing selectively on adjacent evidence where it illuminates the underlying measurement questions.

The paper distinguishes between observed facts, published evidence, analytical interpretation and proposals for future industry work. The priorities developed in Chapter 11 are proposals for consideration and further development; they do not represent agreed industry commitments or imply endorsement by organizations or individuals consulted during the research. The strength and granularity of any conclusion should be understood as bounded by the underlying data, methodology, assumptions and research design.

Chapter 1:

Foundational principles: what attribution can – and cannot – tell us

The central problem in attribution is simple to state and difficult to solve: observing what happened **after** advertising is not the same as knowing what happened **because** of advertising.

Almost every other measurement question follows from that distinction. Identity determines which records can be connected. Attribution rules determine how observed outcomes are allocated. Experiments and causal models attempt to estimate what would have happened otherwise. Consumer journeys, prior purchase propensity, creative variation and time all affect how confidently those conclusions can be drawn.

Eleven foundational principles establish the boundaries within which credible outcomes claims should be interpreted.

Exhibit: The foundational principles of attribution and outcomes measurement

Foundational principle	What it means
1. Outcomes exist at different levels	Advertising can affect exposure, attention, memory, attitudes, behavior and ultimately commercial value. These outcomes are related, but they are not interchangeable, and effects may emerge over different time horizons.
2. Attribution is not the same as incrementality	Attribution assigns credit for an observed outcome. Incrementality asks the causal question: would the outcome have occurred without the advertising? That requires an estimate of the counterfactual.
3. A consumer journey is an observed sequence, not necessarily a causal chain	Touchpoint data can describe what happened before a conversion, but it does not establish that each exposure caused the next event or contributed to the final outcome. Important influences may also be unobserved.
4. Exposed and unexposed consumers may differ before advertising occurs	Advertising exposure is often correlated with consumers' underlying propensity to respond. Targeting, media consumption, demographics, geography and market conditions can influence both exposure and outcomes.
5. Advertising response is heterogeneous	Different consumers can respond very differently to the same advertising. New, lapsed and loyal customers, for example, may have materially different incremental responses even when their observed conversion rates are similar.
6. Advertising effects are dynamic, cumulative and nonlinear	Advertising effects can persist, accumulate and decay over time. Additional frequency often produces diminishing returns, while channels may reinforce, complement or substitute for one another.
7. Exposure and outcome data may relate to different units	Advertising and outcomes can be observed at the person, device, browser, account, household or geographic level. Identity resolution can connect these units, but a successful data match does not necessarily demonstrate individual exposure or causality.

Foundational principle	What it means
8. Randomized experiments provide the strongest design-based causal evidence where validly implemented	Randomization provides the strongest basis for estimating whether a defined intervention caused an incremental outcome. Its bounded estimand should be distinguished from threats to validity: an experiment does not need to reconstruct every pathway or mechanism to answer the causal question it was designed to test.
9. Different methods answer different questions	Attribution, experiments, quasi-experiments, MMM, brand research and customer analytics operate at different levels and illuminate different dimensions of effectiveness. No single methodology provides a complete view.
10. Precision, confidence and decision usefulness are different things	Highly granular outputs can appear precise without being causally reliable. A channel-, household- or impression-level ROAS estimate may contain substantial uncertainty arising from identity error, missing data, selection bias and modeling assumptions.
11. Advertising exposure is not a homogeneous treatment	Two exposures to the same brand are not necessarily equivalent. Creative message, branding, imagery, narrative, offer, format and execution can materially affect attention, memory, persuasion and response. Where creative differs systematically across audiences, channels or platforms, measured performance differences may reflect creative variation as well as media or audience effects.

1. Outcomes exist at different levels

Advertising can produce outcomes at several levels:

Exposure → attention → memory → attitudes → behavior → commercial value

These outcomes are related but not interchangeable. An advertisement can increase awareness without producing an immediate sale, stimulate search without securing the eventual conversion, reinforce an existing customer relationship rather than recruit a new buyer, or generate an immediate response without creating lasting brand effects.

The outcome selected for a study should reflect both the business question being asked and the plausible way in which the advertising could affect behavior. For example, awareness and memory may be appropriate near-term measures for a brand-building campaign, while inquiries, purchases or renewals may be more relevant for advertising intended to prompt an immediate action. Where possible, measurement should also examine whether these intermediate effects ultimately contribute to incremental commercial value.

Outcomes can be measured at several levels:

- **Media outcomes:** reach, frequency, completed views and attention.
- **Intermediate outcomes:** website visits, searches, inquiries, store visits or product-page views.
- **Brand outcomes:** awareness, consideration, preference, trust or purchase intent.
- **Behavioral outcomes:** purchases, subscriptions, registrations, renewals or churn reduction.
- **Economic outcomes:** incremental revenue, contribution margin, customer lifetime value or profit.

A measurable event is not necessarily the most important outcome. Nor is the outcome closest to an advertising exposure necessarily the one most affected by it.

Implications

Every study should state explicitly which outcome is being measured, over what period, and why it is relevant to assessing advertising effectiveness. A platform-reported conversion, a retailer sale and a change in long-term customer value capture different dimensions and time horizons of effectiveness. They may provide different answers to the same overarching question and should not be treated as interchangeable.

2. Attribution is not the same as incrementality

Attribution seeks to assign credit for an observed outcome. Incrementality asks whether the advertising changed the outcome.

Consider a customer who sees advertising and subsequently buys, in which case:

Advertising exposure → purchase

An attribution system may connect the two events and assign some or all of the purchase to the advertisement. But the causal question is different: Would this customer have purchased if the advertising exposure had not occurred?

This counterfactual cannot be observed directly for the same person at the same time. It must be estimated using a randomized control group, a credible quasi-experiment or a model capable of constructing an appropriate comparison.

In simplified notation:

$$\text{Incremental effect} = Y(1) - Y(0)$$

At the individual level, this represents the difference between two potential outcomes: what would happen if the individual were exposed to the advertising and what would happen if the same individual were not. Because both potential outcomes cannot be observed for the same individual under the same conditions, experiments and other causal designs generally estimate population-level quantities such as average treatment effects or intent-to-treat effects rather than directly observing an individual causal effect.

This distinction is particularly important when advertising reaches people **already** likely to buy. A loyal customer may see a search advertisement immediately before making a habitual purchase:

Existing preference → brand search → paid-search click → purchase

The click is attributable in the sense that it appears in the observed journey, but the purchase may not be incremental if the customer would otherwise have clicked the organic result, visited the site directly or bought through another channel.

Large-scale field experiments at eBay found precisely this problem: conventional observational estimates substantially overstated paid-search returns because existing and frequent customers were both more likely to click and more likely to purchase, irrespective of being exposed to the relevant advertising. Effects were materially stronger among new and infrequent customers².

Implications

Attribution can describe pathways, allocate operational credit or support optimization without establishing incremental advertising value.

2 Source: Thomas Blake, Chris Nosko and Steven Tadelis, *Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment* (2015).

3. A consumer journey is an observed sequence, not necessarily a causal chain

Contemporary consumer journeys can include many advertising, information and distribution environments:

TV → CTV → Meta → TikTok → Google → Amazon → retailer website → physical store

This notation describes a possible sequence of observed contacts, but it does not demonstrate that each contact caused the next one or that every contact contributed to the final sale.

In this example, several different processes may be present:

- The TV exposure may create awareness that later stimulates search.
- Social content may reinforce or weaken an existing preference.
- Search may capture demand created elsewhere.
- Amazon or a retailer may influence the final brand decision through price, availability, reviews or merchandising.
- The store visit may have occurred independently of all identifiable digital contacts.
- Some apparently relevant exposures may have had no effect at all.

However, if measurement is the goal, the observable journey is rarely complete – for example, it might include the following components:

Observed: CTV → Google search → purchase

Unobserved: word of mouth → competitor exposure → prior product experience

→ price promotion → in-store availability

Expanding the number of observable touchpoints may provide a fuller account of the consumer's contacts and reduce reliance on any single interaction, but it does not by itself establish which touchpoints caused the outcome. Important influences may remain unobserved, and the additional touchpoints may reflect the consumer's pre-existing propensity to buy rather than effects of the advertising. More complete path data can strengthen attribution analysis, but causal conclusions still require credible counterfactual evidence.

Implications

Consumer-path data should be interpreted as evidence about sequence and association.
Causal credit requires additional assumptions or experimental evidence.

4. Differences between exposed and unexposed consumers can distort attribution

People exposed to advertising can differ systematically from those who are not exposed. Platforms optimize delivery toward users expected to respond. Addressable campaigns focus on certain audiences and exclude others. Heavy category buyers may consume different media from light buyers. Loyal customers may search for a brand more frequently. Retail media is often delivered close to an already likely purchase.

The underlying relationship may look like this:

Prior propensity → probability of advertising exposure

Prior propensity → probability of purchase

Prior propensity can be a confounder when it affects both the probability of advertising exposure and the probability of the outcome. For example, consumers already likely to buy may be more likely to receive advertising because of targeting, while also being more likely to purchase regardless of that advertising. If this prior propensity is not adequately accounted for, the model may mistake pre-existing demand for advertising impact.

When both relationships are present, exposed and unexposed consumers may differ in their underlying likelihood of purchasing. If this difference is not adequately accounted for, the model may mistake pre-existing demand for advertising impact. If advertising exposure is randomized, prior propensity may still predict purchase, but it should not systematically predict exposure and should not confound the estimated advertising effect.

This problem can be especially pronounced in addressable advertising. Targeting criteria and platform delivery algorithms may make consumers with a higher prior propensity to respond more likely to receive the advertising. In such cases, observational methods must distinguish the effect of advertising from the pre-existing characteristics that influenced both exposure and response. However, the problem is not confined to addressable campaigns. Exposure to non-addressable advertising may also be systematically related to purchase propensity because media consumption, geography, demographics and market conditions influence both who encounters a campaign and who is likely to buy. The relevant issue is whether exposure was independent of consumers' underlying propensity to produce the measured outcome.

Research comparing observational estimates with randomized experiments demonstrates how difficult this can be. One influential study examined 15 large Facebook advertising experiments. For each campaign, the researchers compared the causal effect measured through random assignment with estimates produced by commonly used observational methods, including matching and regression adjustments using extensive demographic and behavioral data.

The observational estimates frequently differed materially from the experimental results. They generally overstated advertising effectiveness, although they understated it in some cases. For purchase outcomes, the observational estimate differed from the experimental estimate by a factor of three or more in half of the studies. The finding was not that observational measurement invariably fails, but that rich consumer-level data and sophisticated statistical adjustments may still be unable to remove selection bias. What is missing may not be additional data or computing power, but variation in exposure that is independent of consumers' existing propensity to respond – the function performed by random assignment in an experiment³.

Subsequent research by some of the same authors examined the problem on a much larger scale. Drawing on 663 Facebook experiments and more than 5,000 consumer-level variables, the researchers assessed whether newer observational and machine-learning methods could reproduce the causal effects measured through random assignment. Some methods performed better than others, but none did so reliably across the experiments. Together, the studies provide strong evidence that, in addressable environments where targeting and delivery may be related to consumers' prior propensity, greater data granularity does not necessarily identify or eliminate all the factors that jointly influence exposure and response.

The same underlying risk can arise in non-addressable media, although through less direct mechanisms. Media consumption, geography, demographics and market conditions may influence both the probability of exposure and the probability of purchasing. The critical question is whether exposure is independent of consumers' underlying propensity to produce the measured outcome⁴.

Implications

Studies should examine whether underlying propensities to produce the measured outcome also influenced their likelihood of exposure. This risk is especially pronounced in addressable advertising, where targeting and delivery algorithms may concentrate advertising among consumers already likely to respond, but it can also arise in non-addressable media through differences in media consumption, demographics, geography and market conditions. Granular data and statistical adjustment can reduce observable differences between exposed and unexposed groups but cannot guarantee that selection bias has been eliminated. Where credible causal estimates are required, observational analysis should be validated or calibrated against randomized or well-designed quasi-experimental evidence wherever feasible.

³ Source: Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava and Dan Chapsky, *A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook* (2019).

⁴ Source: Brett R. Gordon, Robert Moakler and Florian Zettelmeyer, *Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement* (2023).

5. Advertising response is heterogeneous

An average advertising effect can conceal very different responses among consumers. Effects may vary according to:

- Prior purchase history.
- Existing brand preference.
- Category involvement.
- Customer tenure.
- Price sensitivity.
- Demographic characteristics.
- Media consumption.
- Competitive exposure.
- Stage in the purchase journey.
- Whether the consumer was already in-market.

A campaign might have the following pattern:

New customer + advertising → Meaningful lift.

Lapsed customer + advertising → Moderate lift.

Loyal customer + advertising → Little incremental lift.

If loyal customers are disproportionately exposed, a naïve attribution study may report very strong performance because it observes many purchases following exposure. An experiment may reveal that much of this purchasing would have happened anyway.

The reverse problem can also occur. Advertising may have a substantial effect on a small strategically valuable segment, but the effect may disappear within the campaign average. Aggregate “return on ad spend” is not a universal property of a channel or platform.

Implications

Robust measurement should account for pre-treatment characteristics – especially previous category and brand behavior – and should report meaningful variation in effects where sample sizes permit. Segmentation developed after observing outcomes should be treated cautiously because it can generate unstable or spurious findings.

The advertising treatment itself can also vary. Two consumers classified as exposed to the same campaign may have received different messages, formats, creative executions or contextual placements. Where those differences vary systematically by audience, publisher or time, a model that treats exposure as a homogeneous treatment can confound variation in media delivery with variation in the advertising itself. Where creative differs materially, measurement should describe and, where feasible, represent the treatment delivered – not simply whether an exposure occurred.

6. Advertising effects are dynamic, cumulative and nonlinear

Advertising does not always operate as a single exposure followed by an immediate response. Effects can accumulate, decay or interact over time:

Advertising at time $t \rightarrow$ Memory at time $t+1 \rightarrow$ Consideration at time $t+2 \rightarrow$ Purchase at time $t+k$.

A current-period sale may reflect recent advertising, accumulated historic advertising, underlying brand strength and other demand drivers. Equally, advertising today may affect outcomes beyond the measurement window.

Three features are especially important.

1. **Carryover.** The effect of advertising can persist after exposure. Short attribution windows may miss delayed effects, while very long windows may capture outcomes that advertising did not cause.
2. **Saturation.** The relationship between exposure and response is not normally linear:

1st exposure > incremental value of 10th exposure

Additional impressions may initially increase response but later produce diminishing returns. This makes spend, impressions, reach and frequency analytically distinct.

3. **Interaction.** Channels may reinforce or substitute for one another:

TV $\rightarrow$ search demand

but also:

paid search $\leftrightarrow$ organic search

A method that measures search in isolation may credit it with demand created by television or social advertising. Conversely, removing paid search may redirect some consumers toward organic search rather than eliminate their purchases.

Modern marketing-mix models attempt to represent these dynamics through adstock, saturation, time-varying coefficients and reach-and-frequency response functions. Bayesian implementations can also represent uncertainty more explicitly and incorporate prior evidence from earlier models, comparable markets or experiments rather than relying on the current historical dataset alone. This can be particularly valuable where media channels move together or the available history contains relatively little independent variation from which to estimate their effects.

Bayesian modeling does not make the analysis assumption-free or causal by construction. Priors can stabilize estimates where the signal is weak, but they can also create apparent precision if they dominate the available evidence. The relevant questions are how the priors were chosen, how strongly the observed data updated them, whether the conclusions survive reasonable alternative assumptions and whether experimental or other external evidence supports the result⁵.

⁵ Source: Chris Williams and Jon Watts, *Models as Masters? Marketing Mix Modeling and AI-Driven Media Decision Making* (July 2026).

Implications

Attribution and outcomes measurement solutions need to account for advertising effects that may accumulate across repeated exposures, persist after exposure and change as frequency increases. Studies that consider only outcomes occurring immediately after an exposure may miss delayed effects, while those that assign equal value to every additional exposure may overstate the contribution of high frequency. Where the evidence permits, analysis should also examine whether channels reinforce one another or merely capture demand generated elsewhere. The timing and allocation of effects should be based on evidence about how consumers respond, not determined solely by a fixed attribution window or a model's default settings.

7. Exposure and outcome data may relate to different levels or units

The advertising exposure and the outcome may be observed at different levels:

Person → device → browser → account → household → geographic area

A CTV exposure may be recorded at the household or device level, while a purchase is recorded against a retailer account or loyalty-card identity. A mobile location signal may suggest a store visit but not identify the purchaser. A television set may be on without establishing which household member saw the advertisement.

Identity resolution attempts to connect these units:

CTV device ID ↔ household ID ↔ retailer identity ↔ purchase

But the resulting connection may be deterministic, probabilistic or inferred. Even a correct household-level match does not necessarily prove person-level exposure. A technically successful data match can create more precision in the record without producing equivalent precision in the causal conclusion.

Coverage is equally important. Measurement may exclude cash purchases, unauthenticated viewing, walled-garden exposures, non-participating retailers or consumers who cannot be matched. The measured population may differ systematically from the population of interest.

Implications

Every attribution result should identify the exposure unit, outcome unit, matching method, match rate, excluded population and likely direction of bias. "Deterministic" should describe the identity link, not the causal inference.

8. Randomized experiments provide the strongest design-based causal evidence where they can be validly implemented

Where random assignment is feasible and treatment is implemented as intended, randomized experiments generally provide the strongest design-based basis for estimating the causal effect of a defined intervention:

Random assignment → Eligible for advertising → Outcome

Random assignment → Held out from advertising → Outcome

Incremental effect = Difference in outcomes between the two groups

Randomization is powerful because, in expectation, it makes treatment and control groups comparable not only on observed characteristics but also on unobserved characteristics that might otherwise influence both advertising exposure and the outcome. Differences in subsequent outcomes can be interpreted causally, subject to the validity of the experimental design and implementation.

Randomized experiments can operate at different units depending on the advertising environment and the question being tested. Treatment can be assigned at the individual, household, device, store, geographic or market level. The appropriate unit should reflect both the way advertising can be varied and the level at which outcomes can be measured credibly.

Geo-experiments are particularly useful where user- or household-level randomization is unavailable or where the advertiser wants to estimate the effect of a broader change in media activity. Geographic units are assigned to treatment and control conditions, and the analysis estimates how outcomes changed relative to a credible counterfactual. Google's work on geo-experiments and Bayesian structural time-series models illustrates both the potential of these approaches and the importance of careful design and counterfactual construction.

Randomized experiments are strongest when they answer a bounded, decision-relevant causal question. That boundedness is a property of the **estimand** – the quantity the experiment is designed to estimate – rather than a weakness of the methodology. If an advertiser randomly varies whether a defined campaign, channel or increase in media investment is available to otherwise comparable populations, the experiment can estimate whether that intervention caused the specified outcome to change. It does not need to reconstruct every intermediate step in the consumer journey to answer that question credibly.

Two different kinds of boundary should be distinguished.

The first concerns **threats to experimental validity**. Insufficient statistical power, failed or inappropriate randomization, weak treatment contrast, non-compliance, treatment leakage, spillovers or contamination between treatment and control groups, and material problems in outcome measurement can weaken whether an experiment supports the causal claim it was designed to estimate. These are genuine limitations of the evidence and should form part of the assessment of experimental quality.

The second concerns **the scope of the estimand and the generalizability of the result**. A campaign-level experiment may establish whether the campaign generated incremental sales without identifying the separate causal contribution of every publisher, impression or creative execution. It may not explain the mechanism through which the effect occurred, capture effects beyond the observation period or establish that the same result would occur at a different spending level, for a different audience or under different market conditions.

These are important boundaries on interpretation, but they are not failures of causal identification. They are different questions unless the experiment has been designed specifically to answer them.

For example, an experiment that randomizes a campaign package may credibly establish:

Campaign package → Incremental sales

without establishing:

TV=30%; CTV=20%; Search=25%; Retail media=25%

The experiment should not be judged deficient because it cannot support a causal decomposition that its design did not create. If the advertiser needs causal evidence about the incremental contribution of an individual component, the relevant intervention must itself be varied in a way that supports that comparison.

The same distinction applies to creative. A campaign holdout can estimate the incremental effect of the campaign as implemented, including its creative strategy. Random assignment among defined creative executions can estimate differences between those executions. Observational differences in outcomes among creatives may remain useful, but they should not inherit causal status merely because another element of the campaign was randomized.

For causal questions that can be randomized credibly, randomized evidence should receive greater weight than observational association. **Methodological pluralism should not imply methodological equivalence.** Attribution, MMM, brand research and customer analytics can extend coverage, examine longer horizons, describe pathways, investigate mechanisms and produce more granular hypotheses. But when the claim is causal, the strength of the conclusion should reflect the identification strategy supporting it.

Implications

Randomized experiments should be used where the causal question can be defined clearly, treatment can be varied credibly and the decision warrants the required cost and operational complexity. Their value should be judged against the causal question they were designed to answer, not against whether they provide a complete explanation of every exposure, mechanism or consumer pathway associated with the campaign.

Experimental design and reporting should state clearly what was randomized; the treatment and reference conditions; the unit of randomization; the population and outcome being studied; the treatment and observation periods; and the estimand being reported. Where relevant, researchers should distinguish assignment from actual exposure and identify whether the estimate is an intent-to-treat effect or another treatment-effect quantity. Statistical uncertainty, power, non-compliance, spillovers, treatment contamination, outcome coverage and material limits to generalization should also be disclosed.

The granularity of the causal claim should remain aligned with the level at which causal variation was created. A campaign-level randomization can support a campaign-level causal conclusion; it does not automatically establish the incremental contribution of individual publishers, audiences, creative executions or impressions within the campaign. More granular attribution and observational analysis can provide useful diagnostic evidence at those levels, but stronger causal claims require correspondingly stronger identification.

Experiments should operate as causal anchors within a broader evidence system. Experimental findings can test, calibrate or constrain attribution and MMM; attribution can provide more granular evidence about observable pathways; MMM can place bounded experimental results within a broader cross-channel and longer-term context; and brand and customer research can illuminate mechanisms and outcomes not captured in immediate transactional data. Where these sources differ, evidence with stronger causal identification should generally receive greater weight for causal claims.

9. Different methods answer different questions

There is no single measurement system capable of observing every relevant exposure, constructing every counterfactual and estimating both immediate and long-term effects at arbitrary levels of granularity. Given this, the strongest measurement frameworks combine complementary methods.

Exhibit: Complementary outcomes-measurement methodologies

Method	Principal strength	Principal limitation
Path-based attribution	Describes observable journeys at granular level	Often confuses sequence with causality
Randomized experiments	Strong causal identification	Usually local to a campaign, platform, population or period
Quasi-experiments	Can estimate causality where randomization is unavailable	Depend on demanding identification assumptions
Marketing-mix modeling	Covers multiple channels and longer periods	Aggregate, specification-sensitive and dependent on sufficient variation
Brand-lift studies	Measures intermediate attitudinal effects	Survey response and linkage to commercial outcomes can be difficult
Customer and retailer analytics	Provides granular behavioral outcomes	Identity, coverage and selection limitations remain
Qualitative research	Explains mechanisms and consumer interpretation	Does not quantify population-level incrementality

These methods don't produce identical answers, but can be used to test and calibrate one another:

Experiments + MMM + attribution + brand research → triangulated evidence

The relationship between experiments and MMM is especially important. MMM can examine the total business, multiple media channels and relatively long periods, but normally learns from observational variation in historical activity. Randomized and well-designed quasi-experiments can create stronger counterfactual evidence, but usually test a much more bounded intervention: a particular channel, campaign, geography, audience, spending change or period.

Used together, each can strengthen the other. Experimental lift estimates can be used to calibrate or constrain an MMM, helping the model distinguish effects that historical time-series variation alone may struggle to separate. The uncertainty and scope of the experiment should carry into that calibration rather than being treated as fixed ground truth. Conversely, an MMM can identify channels, spending ranges or decisions in which uncertainty is commercially important and where additional experimentation would be most valuable.

The objective is a continuing learning system rather than a contest between methodologies: MMM identifies important uncertainties; experiments create stronger evidence in selected areas; the resulting findings update or challenge the model; and the revised model informs subsequent allocation and testing decisions. Attribution data can add more granular evidence about observable pathways, while brand and consumer research can capture mechanisms and longer-term effects that may not yet be visible in transactions⁶.

6 Ibid.

Implications

No single method should be expected to provide a complete account of advertising effectiveness. Path-based attribution, experiments, marketing-mix modeling, brand research and customer analytics answer different even if often related questions, operate at different levels of granularity and have different strengths and limitations. The appropriate combination will depend on the decision being made and the complexity of the consumer journey.

Combining methods does not mean averaging their results or assuming that agreement proves accuracy. Researchers must first determine whether the methods are measuring comparable interventions, outcomes, populations and time periods. Apparent differences may reflect genuine differences in scope – for example, an experiment may estimate the short-term incremental effect of a particular campaign, while an MMM estimates the contribution of a channel over a longer period and an attribution system describes the observable paths preceding conversion.

The objective is to assess whether the findings can be reconciled into a credible explanation of what happened, what would probably have happened without the advertising and how confidently the effect can be attributed at the level claimed. Where results conflict, the discrepancy should be investigated rather than concealed: it may reveal differences in definitions, missing data, identity limitations, model assumptions, time horizons or selection bias. Evidence from methods with stronger causal identification should generally carry greater weight for causal claims, while more granular observational methods may remain valuable for describing journeys and generating hypotheses.

Triangulation is a structured analytical process, not a simple test of whether several models produce similar numbers. Measurement quality should be judged by whether the conclusions are supported by appropriately chosen and transparently reconciled evidence – and whether the remaining uncertainty and disagreement are clearly disclosed.

10. Precision, confidence and decision usefulness are different things

Attribution systems frequently produce highly granular outputs – down to a channel, placement, creative, household or impression. Granularity should not be confused with accuracy.

A model might report:

CTV campaign = \$4.37 in attributed revenue per \$1.00 spent.

But the true uncertainty may be much wider because of identity errors, incomplete exposure data, selection bias, model specification, organic demand and sampling variation.

A more decision-useful conclusion could be:

Probability that incremental ROAS is greater than 1 = 75%.

or: The evidence indicates that the campaign was incrementally positive, but it does not support reliable allocation of the effect across individual publishers.

This distinction becomes more important as consumer journeys grow more complex. The appropriate ambition may shift from deterministic event-level attribution toward bounded estimates, probabilistic conclusions and structured learning over time.

Implications

Vendors and advertisers should disclose uncertainty, sensitivity to assumptions and the level at which findings remain credible. A less granular causal estimate may be more useful than an extremely granular correlational one.

11. Advertising exposure is not a homogeneous treatment

Outcomes measurement often reduces advertising treatment to a relatively simple distinction:

Exposed → outcome

versus:

Not exposed → outcome

That distinction can be analytically useful, but it can obscure an important source of variation. Two consumers classified as exposed to the same campaign may not have received equivalent advertising treatments. They may have encountered different creative executions, messages, formats, offers, durations, contextual placements or combinations of these elements.

Creative matters because media delivery creates the opportunity for advertising to work, but the creative execution determines what the consumer encounters within that opportunity. Differences in imagery, narrative, branding, sound, product presentation, distinctive assets, calls to action and other executional characteristics can affect whether an advertisement attracts attention, evokes emotion, creates or refreshes memory, communicates the brand, changes perceptions or stimulates behavior⁷.

An impression should not be treated as a standardized unit of advertising effect.

This becomes particularly important where creative varies systematically across audiences, publishers, platforms or periods. Consider two publishers carrying the same campaign:

Publisher A → Creative A → outcome

Publisher B → Creative B → outcome

If Creative A is materially more effective than Creative B, a measurement system that represents only publisher exposure may attribute the resulting performance difference to the media environment even though some of the difference was created by the advertising itself. The same problem can arise where different audience segments receive different creative, where particular executions run predominantly at different stages of a campaign, or where different formats are concentrated in particular channels.

Addressable and dynamically optimized advertising make the issue more consequential. A campaign may simultaneously change who receives advertising and what advertising they receive:

Audience selection + media delivery + creative personalization → outcome

In these circumstances, the advertising treatment is a package rather than a single exposure. If consumers with particular characteristics are systematically served particular creative executions, observed differences in creative performance may reflect differences in the consumers receiving the creative as well as differences caused by the creative itself. Optimization systems can intensify this selection problem when they learn to serve particular executions to consumers predicted to respond.

⁷ Source: ARF Cognition Council Cognitive Processes Working Group, *The Brain on Advertising: Why We Notice, Feel, Remember and Act on Ads* (May 2026).

The research design should reflect the causal question being asked.

If the question is whether one media strategy outperforms another, creative should ideally be held sufficiently comparable across the conditions being evaluated so that differences in outcomes can reasonably be associated with the change in media strategy.

If the question is whether one creative execution outperforms another, random assignment among creative treatments can provide stronger evidence:

Random assignment → Creative A → outcome

versus:

Random assignment → Creative B → outcome

If audience strategy and creative both change, the resulting experiment may credibly estimate the incremental effect of the combined strategy without establishing how much of the effect resulted from audience selection and how much resulted from creative. Multi-arm or factorial experiments can sometimes separate these effects and test interactions among audience, creative and media conditions, although the number of treatment combinations and required sample size can increase rapidly.

The same principle applies to observational attribution and marketing-mix modeling. Where material creative variation is omitted from the analysis, some effects associated with channels, publishers, audiences or periods may partly reflect differences in the creative deployed within them. Creative identifiers, executional metadata, pre-testing, brand-lift evidence and controlled creative experiments can help make this variation visible. But the existence of highly granular creative data does not itself establish creative-level causality. A model may identify creative characteristics associated with stronger outcomes without demonstrating that changing those characteristics would produce the predicted change in performance.

Creative effectiveness is also not necessarily reducible to a single universal score. An execution may attract attention without strengthening brand memory; another may generate relatively little immediate response while contributing to longer-term familiarity, distinctiveness or consideration. The relevant measure depends on the advertising objective, audience, context, time horizon and outcome being assessed.

Implications

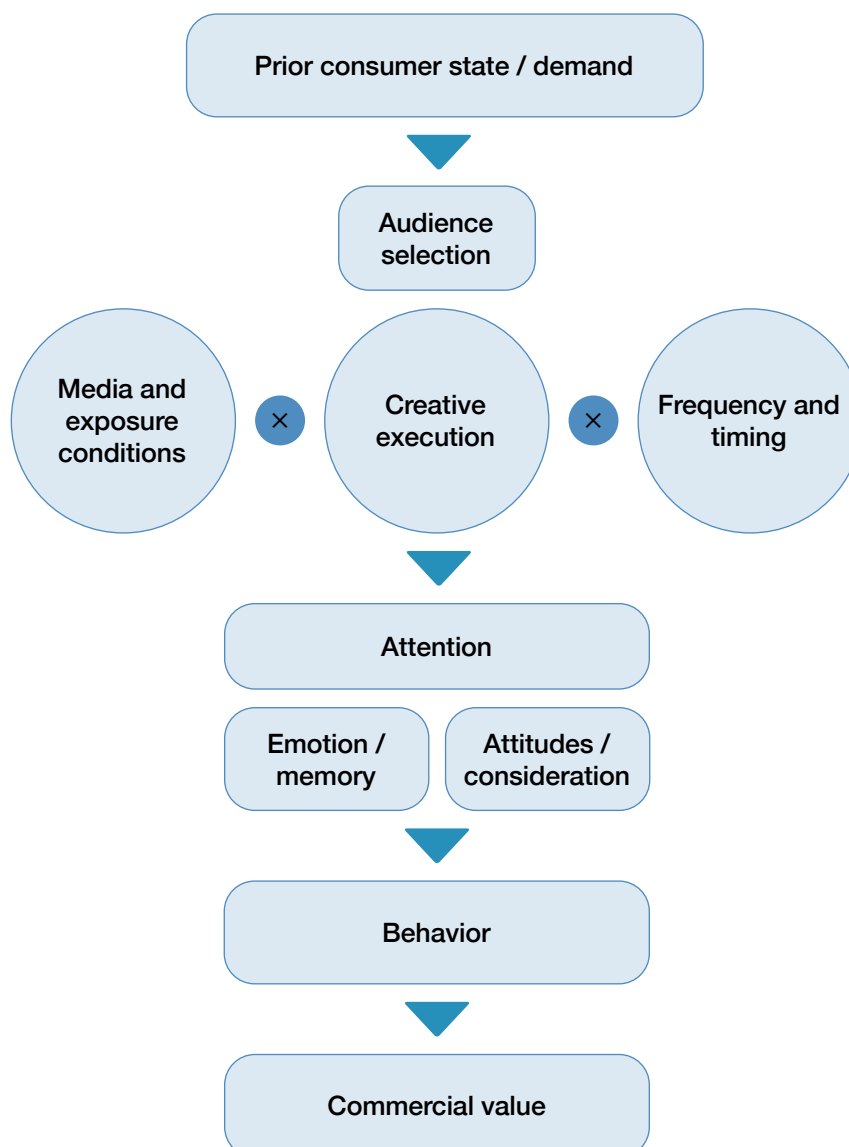
Attribution and outcomes studies should identify material variation in the advertising treatment rather than record exposure as though every impression were equivalent. Where relevant, researchers should state which creative executions were included, whether creative assignment was fixed, randomized, targeted or optimized, and whether creative differed systematically across audiences, publishers, platforms or treatment conditions.

The interpretation of the result should then match the research design. A campaign-level experiment may estimate the incremental effect of the campaign as implemented, including its creative strategy, without identifying the separate contribution of individual executions. A randomized creative test may support a causal comparison between defined executions. An observational creative-performance report may reveal useful associations without establishing that the creative itself caused the difference.

Measurement should distinguish, where the evidence permits, the effects of creative execution from the effects of audience selection, media delivery and context – and should describe the result as the effect of a combined advertising treatment where those elements cannot credibly be separated.

Media supplies the opportunity; creative supplies the stimulus; the consumer supplies the starting state; measurement tries to identify what changed.

Exhibit: Advertising effectiveness is produced by a system, not by exposure alone



Conclusion: Attribution is a question of degree, not a binary capability

Outcomes can be observed at different levels; attribution and incrementality answer different questions; consumer pathways are not necessarily causal chains; exposure can be related to prior purchase propensity; advertising effects vary across people, time, media and creative treatments; and greater output granularity does not necessarily imply greater causal confidence.

Exhibit: What the foundational principles mean for attribution and outcomes measurement

Foundational principle	Implication for attribution and outcomes measurement
1. Outcomes exist at different levels	Measurement should specify the outcome being assessed and use metrics appropriate to the campaign objective and time horizon.
2. Attribution is not incrementality	An observed conversion following exposure should not automatically be interpreted as a conversion caused by advertising.
3. Consumer journeys are not necessarily causal chains	Touchpoint data can describe the path to an outcome, but causal credit requires additional evidence.
4. Exposure is often correlated with prior propensity	Measurement must distinguish advertising effects from consumers' pre-existing likelihood to buy, especially where targeting is based on behavioral or purchase data.
5. Advertising response is heterogeneous	Average campaign effects may conceal important differences between new, existing, lapsed and other customer groups.
6. Advertising effects are dynamic and nonlinear	Measurement should account for carryover, frequency, diminishing returns and interactions between channels rather than rely solely on fixed attribution windows.
7. Identity and exposure are imperfectly observed	The reliability of attribution depends on the quality of identity matching, exposure measurement and the alignment between the unit exposed and the unit producing the outcome.
8. Randomized experiments provide the strongest design-based causal evidence	Use randomized assignment where feasible for causal questions. Distinguish threats to experimental validity from limits on the scope or generalizability of the estimand, and keep the causal claim aligned with what was actually randomized.
9. Different methods answer different questions	Attribution, experiments, MMM, brand measurement and other approaches should be combined as complementary sources of evidence rather than treated as competing universal solutions.
10. Precision is not the same as certainty	Measurement should communicate uncertainty and avoid presenting highly granular estimates with greater confidence than the underlying data and methodology justify.
11. Advertising exposure is not a homogeneous treatment	Measurement should identify material creative variation and distinguish, where the research design permits, the effect of creative from the effects of audience selection, media delivery and context.

The implication is that an outcomes result should be judged not by the apparent precision of the number alone, but by the claim being made and the evidence supporting that claim. The next question is more precise: what exactly was measured, and what is being inferred from it?

Chapter 2:

What exactly are we measuring and what are we inferring?

The same purchase can support several very different measurement claims.

A system may observe that the purchase followed an advertising exposure. An attribution model may assign credit to one or more preceding touchpoints. A causal analysis may estimate that the purchase would not have occurred without the advertising.

These claims can refer to exactly the same transaction. They do not mean the same thing.

Outcomes measurement often collapses two different questions: what was measured, and what is being inferred from it. A purchase can be observed, associated with advertising, attributed to particular touchpoints or estimated to have been caused by a defined advertising intervention. Each represents a different evidentiary claim.

Greater data granularity, identity resolution and model sophistication can make the observable path increasingly precise without necessarily strengthening the causal inference. A highly detailed account of what happened before a conversion is not the same as evidence of what would have happened without the advertising.

Two dimensions should remain distinct:

1. **What was measured:** advertising delivery, a subsequent response, a brand effect, a transaction or an economic outcome.
2. **What was inferred:** a descriptive observation, a statistical association or prediction, an allocation of attribution credit, or a causal/incremental effect.

These dimensions should remain separate. An observed purchase is a business outcome. An attributed purchase is the output of a rule or model assigning credit for that outcome. An incremental purchase is an estimate of a purchase that would not have occurred under the specified counterfactual.

1. Exposure: did the advertising opportunity occur?

At the most fundamental level, measurement asks whether advertising was delivered:

- Was an impression served?
- Was the advertisement viewable?
- Did it appear on the television screen?
- Was a household reached?
- How many times was it reached?
- Was the advertisement completed or attended to?

These are primarily **media-delivery questions**. They tell us something about the opportunity for advertising to have an effect, but not whether an effect occurred.

Even here, apparently simple statements require care. A served impression is not necessarily a viewable impression. A television set displaying an advertisement does not prove that a particular household member watched it. Attention measures attempt to move closer to actual human exposure but remain measures of contact or engagement rather than proof of subsequent advertising effect.

Exposure is a critical input for many forms of outcomes measurement, but it is not itself an outcome in the commercial sense. At this first level, the measurement question is simply **whether there was an opportunity for the advertising to affect the consumer**.

2. Response: what happened after the exposure?

The next level of measurement sets out to determine whether some action or change occurred following exposure to the advertisement.

This might include immediate behavioral responses:

Exposure → click

Exposure → website visit

Exposure → search

Exposure → store visit

Exposure → purchase

Responses don't necessarily need to be transactional. Advertising can also be followed by changes in memory, awareness, consideration, preference or purchase intent. Brand-lift research, for example, can measure differences in brand perceptions rather than waiting for an eventual transaction.

This distinction matters because advertising operates through multiple mechanisms and over different time horizons. An advertisement designed to create awareness should not necessarily be judged solely by whether someone purchased within seven days.

Response measurement answers a useful descriptive question: **What happened after exposure to the advertising?** However, temporal sequence alone does not establish that the advertising produced the response.

3. Association: are exposure and outcome statistically related?

Measurement becomes more ambitious when it tests whether advertising exposure and an outcome are systematically and statistically associated.

Suppose purchasers were twice as likely as non-purchasers to have encountered a particular campaign. Or exposed households purchased at a higher rate than apparently unexposed households. This appears to establish an **association**:

Advertising exposure ↔ higher probability of purchase

Association can be highly informative. Many forms of customer analytics, path analysis and predictive modeling depend upon it.

But exposed and unexposed consumers can differ systematically **before** the advertising occurs. People already interested in a product may be more likely **both** to encounter advertising **and** to buy. Targeting algorithms may deliberately target and deliver advertising specifically at people predicted to convert. Loyal customers may be more likely to search for the brand and subsequently encounter paid search.

Put another way:

Exposure associated with purchase ≠ exposure caused purchase

This distinction between association and causation is one of the most important boundaries in outcomes measurement.

4. Attribution: which observed marketing contacts should receive credit?

Attribution asks a different question. It normally starts with an observed outcome – a purchase, registration, subscription or other conversion – and works backwards through the observable journey:

CTV → social → search → retail media → purchase

The attribution problem is: **How should credit for the purchase be attributed to the preceding touchpoints?**

Last-touch attribution assigns the outcome to the final qualifying interaction. Multi-touch approaches distribute credit using rules, statistical models or machine learning. As such, attribution involves assigning credit to digital touchpoints found on the user's path to conversion⁸.

This can be extremely useful operationally. Attribution can identify common paths, compare touchpoints, support bidding systems and help marketers understand how observable interactions relate to conversion.

But attribution is fundamentally a **credit-allocation problem**. Causal inference is a different problem. Conventional attribution starts with an observed conversion and allocates credit across the advertising touchpoints that preceded it. Experimental measurement starts from a different direction: consumers are randomly assigned to treatment and holdout groups, and subsequent outcomes are compared to estimate the incremental effect of the advertising⁹.

The ideal attribution system would allocate credit in proportion to the **incremental contribution** of each touchpoint, but not every attribution methodology succeeds in doing this.

5. Incrementality: how many additional outcomes did the advertising produce?

This is the critical step from observed outcomes toward effectiveness.

Suppose 10,000 consumers exposed to a campaign subsequently purchase.
A conventional attribution report might show:

10,000 attributed conversions

But imagine that, without the advertising, 8,000 of those consumers would have purchased anyway.

The economically relevant result is then:

10,000 observed conversions

– 8,000 counterfactual conversions

= 2,000 incremental conversions

This distinction transforms the interpretation of campaign performance.

The first number answers: **How many conversions can be connected to the advertising under the attribution rules?**

⁸ Source: WARC, *What we know about attribution and marketing mix modelling* (July 2026).

⁹ Source: Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero and German Schnaidt, *Amazon Ads Multi-Touch Attribution* Amazon, 11 August 2025, pp. 3–4.

The second asks: **How many additional conversions occurred because of the advertising?**

Attributed sales based on tracked purchases can substantially exceed the incremental sales generated by advertising¹⁰.

Attribution models need comparison with independent measures of campaign incrementality if their credit allocation is to be evaluated against causal evidence¹¹.

This gives us one of the most important equations in the paper:

Conversions following advertising $\neq$ conversions caused by advertising

or, more formally:

Incremental conversions = observed conversions with advertising – estimated conversions without advertising

The second term is the hard part. It is the **counterfactual**.

6. Causation: did the advertising intervention change the outcome?

Incrementality brings us to the underlying causal question. For an individual consumer, we would ideally observe two parallel worlds:

World A: Consumer sees advertising $\rightarrow$ outcome

World B: Same consumer does not see advertising $\rightarrow$ outcome

The difference would reveal the causal effect. However, both states clearly can't be observed simultaneously, so the counterfactual has to be estimated.

Randomization provides the strongest general solution because treatment and control groups are made probabilistically comparable before the intervention. In an advertising randomized controlled trial (RCT), one group is eligible to receive the advertising while another is held out; differences in subsequent outcomes between the groups can then be used to estimate the incremental causal effect of the advertising¹².

Well-designed quasi-experiments and causal models can also support causal inference where randomization is unavailable, although they generally depend on stronger assumptions about the comparability of exposed and unexposed groups.

Randomization can provide strong evidence that advertising caused an incremental change in the outcome being measured, but the strength of the causal method does not make the resulting claim unlimited in scope. An experiment establishes the causal effect of the particular intervention that was tested, on the particular outcome and population studied.

This distinction becomes especially important in multi-channel campaigns. An RCT demonstrating that an integrated campaign increased sales does not necessarily establish what share of those incremental sales was caused by any single channel – television, CTV, search, social or retail media individually. A campaign-level experiment can provide strong causal evidence about the campaign as a whole while providing much less evidence about the causal contribution of the individual exposures or touchpoints within it.

¹⁰ Source: WARC, *What we know about attribution and marketing mix modelling*, July 2026.

¹¹ Source: Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero and German Schnaidt, *Amazon Ads Multi-Touch Attribution*, Amazon, August 2025, pp. 2–4.

¹² Source: Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero and German Schnaidt, *Amazon Ads Multi-Touch Attribution* (August 2025).

The causal claim might be: **Campaign eligibility caused an estimated 4% incremental increase in purchases among the tested population during the measurement period.**

That is a stronger claim than saying that purchases were associated with advertising exposure but it may also be a much narrower claim than an impression-level attribution system appears to make, assigning a specific outcome to a specific exposure.

These distinctions provide a hierarchy of outcomes-measurement claims. Each successive level makes a stronger inference about the relationship between advertising and an observed outcome – from whether an advertising opportunity occurred, through what happened afterwards and how credit was assigned, to whether changing the advertising caused an incremental change in the outcome.

Incrementality belongs within causal inference rather than representing a separate category beneath causation. An incremental outcome is an estimate of the difference caused by a defined intervention relative to a counterfactual. The relevant distinction is between observing or allocating outcomes and estimating how outcomes changed because the advertising changed.

Moving up this hierarchy requires progressively stronger evidence. Greater causal confidence can nevertheless come at the expense of granularity: an experiment may establish the incremental effect of a campaign more convincingly than an attribution model while providing less information about the contribution of individual exposures within it.

Exhibit: A hierarchy of outcomes-measurement claims

Level	What is being measured or inferred?	Question answered	Example finding	What it does not establish
Exposure	Delivery/contact	Did the advertising opportunity occur?	2 million households reached	That advertising changed behavior
Response	Subsequent behavior or attitude	What happened after exposure?	20,000 exposed users visited the site	That advertising caused those visits
Association / prediction	Statistical relationship	Are exposure and outcome related, or does one help predict the other?	Exposed households purchased more often	That advertising produced the difference
Attribution	Allocation of observed outcomes	Which recorded touchpoints receive credit?	Search receives 40% of conversion credit	That search caused 40% of incremental sales
Causal / incremental estimate	Counterfactual effect of a defined intervention	How did the outcome change because the advertising changed?	Campaign generated 2,000 incremental purchases and increased purchase probability by 4%	Necessarily which individual touchpoints produced the effect, or that the result generalizes beyond the intervention, population, outcome and period studied

This is a hierarchy of claims, not a ranking of methodologies. A causal or incremental claim can be supported by evidence of very different strength. The credibility of the claim depends separately on the research design, counterfactual and identification strategy used to distinguish the advertising effect from other influences on the outcome, implying a hierarchy of evidence for causal claims.

The hierarchy does not imply that lower-level claims are unimportant. Exposure, response, association, attribution and causal or incremental estimation answer different operational and evidentiary questions. The problem arises when evidence established at one level is presented as though it supported a stronger claim at another.

The conversion trap

The term **conversion** deserves particular attention because it is used so extensively in digital and outcomes measurement. A conversion is simply a defined event. Depending on the advertiser, it might mean:

click → site visit → lead → application → subscription → purchase → renewal

The fact that a measurement platform labels something a “conversion” says nothing in and of itself about whether the advertising actually caused it.

Consider a loyal customer who routinely purchases the same product every month. Shortly before the next purchase, the customer encounters a retargeting advertisement, clicks it and buys.

The journey is perfectly observable:

Retargeting impression → click → purchase

The purchase may legitimately qualify as an **attributed conversion** under the measurement rules, but the relevant effectiveness question remains: **Would the purchase have happened anyway?**

If the answer is yes, incremental conversions from that customer equal zero. This explains why conversion optimization and advertising effectiveness are related but not identical objectives. A predictive algorithm may become extremely good at finding people who will convert, but unless the system distinguishes **persuadable consumers from consumers who were going to convert anyway**, improved conversion prediction does not necessarily imply improved incremental advertising effectiveness.

This issue becomes especially acute in the relatively simple journeys identified in Chapter 3. When advertising is delivered close to purchase, outcomes are highly observable and identity resolution is strong, the resulting measurement can *look* exceptionally definitive. Yet these are often precisely the circumstances in which advertising is closest to existing demand.

The better the system becomes at finding likely buyers, the greater the need to distinguish **prediction of conversion from causation of conversion**.

Attributed conversions and incremental conversions answer different business questions

Attributed conversions can support operational questions about where observable response occurs, which campaigns or touchpoints receive credit and how a system should optimize against conversion probability.

Incremental conversions answer a different economic question: **how many outcomes changed because the advertising intervention occurred?**

The distinction carries directly into return measures:

$$\text{Attributed ROAS} = \text{attributed revenue} \div \text{advertising spend}$$

$$\text{Incremental ROAS} = \text{incremental revenue} \div \text{advertising spend}$$

Where the necessary data are available, the economically relevant measure can go further by incorporating contribution margin and incremental marketing cost. A campaign can generate many attributed conversions because it reaches consumers already likely to buy while creating relatively few additional outcomes. The distinction can materially affect investment decisions.

Incrementality does not mean short-term conversion

Incrementality describes the causal logic of a claim; it does not determine which outcome matters. An incremental effect can be estimated for awareness, consideration, leads, purchases, retention, customer value or other outcomes, provided the intervention, outcome and counterfactual are defined appropriately.

A narrow focus on immediate conversion can miss advertising effects that develop through memory, brand preference, delayed purchasing or customer value. Chapter 8 examines these longer-term effects in detail. For present purposes, the important distinction is simply that **the position of an outcome in the consumer or commercial funnel is separate from the evidentiary strength of the claim made about it.**

Every outcomes claim has three dimensions

Every outcomes claim should be understood across **three separate dimensions**.

1. What was measured?

The first dimension is the object of measurement: the exposure, consumer response or business outcome that has actually been observed or estimated. Outcomes can sit at different points in the advertising-response process:

Exposure → attention → memory → attitudes → behavior → commercial value

A measure of awareness, a purchase and customer lifetime value are different measurement objects. Their position in the response or commercial chain does not, by itself, determine how strongly advertising's contribution to them has been established.

2. What is being claimed?

The second dimension is the inference being made about advertising's relationship to the measured result. The same underlying outcome can support claims of different strength:

Observation → association / prediction → attribution → causal / incremental effect

For example, a purchase can be observed; exposure and purchase can be statistically associated; attribution credit for the purchase can be assigned to preceding advertising; or the analysis can estimate how purchase probability changed because a defined advertising intervention occurred.

These are different claims, even when they refer to the same transaction.

3. How strong is the evidence supporting the claim?

The third dimension is evidentiary strength: how credibly the research design distinguishes the effect of advertising from the other factors that could have produced the outcome.

A result described as causal or incremental is not automatically supported by strong causal evidence merely because it carries that label. Randomized designs generally provide the strongest design-based basis for causal inference where they are validly implemented. Credible quasi-experimental approaches can support causal conclusions under additional identification assumptions. Observational attribution and predictive models can provide valuable evidence about pathways, associations, heterogeneity and likely response without necessarily identifying what advertising caused.

These three dimensions should remain distinct:

What was measured is not the same as what is being claimed, and what is being claimed is not the same as how strongly the evidence supports that claim.

A commercially important outcome may be supported only by observational evidence. A highly granular attribution estimate may allocate credit precisely without identifying an equally granular causal effect. Conversely, a randomized experiment may provide strong causal evidence about a relatively narrow outcome, campaign or population without explaining the contribution of every individual exposure.

Any outcomes result should be interpreted through three questions: **What was measured? What is being claimed about advertising's relationship to it? And how strongly does the available evidence support that claim?**

The exhibit below maps the first two dimensions: what is being measured and what is being claimed. The third dimension – how strongly the evidence supports that claim – cannot be represented simply by the position of a

result in the matrix. It depends on the research design, counterfactual, identification strategy, assumptions and implementation supporting the individual analysis.

Exhibit: Outcome level and inferential status are separate dimensions

	Observed	Associated	Attributed	Causal / incremental claim
Exposure / attention	✓	✓	–	Sometimes
Memory / awareness	✓	✓	Sometimes	✓
Consideration / intent	✓	✓	Sometimes	✓
Behavioral response	✓	✓	✓	✓
Conversion / sale	✓	✓	✓	✓
Revenue / profit / LTV	✓	✓	✓	✓ but increasingly difficult

Moving toward more commercially consequential outcomes can increase the practical value of the result, while moving toward stronger causal claims increases the evidentiary burden. These are separate movements. The difficulty of supporting either at the required level of confidence depends on the available data, research design and consumer journey.

Implications

The central requirement for a more mature, transparent outcomes-measurement marketplace is **claim discipline**, distinguishing between these different claims.

Exhibit: What is being measured?

Measurement object	Principal question
Exposure and delivery	Was there an opportunity for the advertising to be seen?
Intermediate and brand response	Did attention, search, visitation, memory, awareness, consideration or preference change?
Behavioral outcome	Did a purchase, subscription, renewal, visit or other action occur?
Economic outcome	What revenue, margin, customer value or profit resulted?

Exhibit: What is being inferred?

Inferential status	What the claim establishes
Observation	An exposure, response or outcome was recorded.
Association or prediction	Exposure and outcome vary together, or one helps predict the other.
Attribution	Credit has been allocated to one or more observed marketing contacts under a defined rule or model.
Causal / incremental estimate	The analysis estimates how an outcome changed relative to a defined counterfactual.

Exhibit: How strongly is the claim supported?

Evidentiary basis	What it contributes
Descriptive observation	Establishes that an exposure, response or outcome was recorded within the measurement boundary.
Observational / predictive analysis	Establishes relationships, patterns or predictive information, but may remain vulnerable to selection and unobserved confounding.
Credible quasi-experimental evidence	Can support causal inference where the assumptions required to construct the counterfactual are credible and adequately tested.
Randomized evidence	Generally provides the strongest design-based evidence that a defined intervention caused an outcome to change, subject to study quality and the bounds of the estimand.

Incrementality belongs within causal inference rather than forming a separate rung below causation. The key question is not whether a result carries the label “**attribution**” or “**incrementality**,” but what claim is actually being made and what counterfactual, research design and assumptions support it.

The hierarchy of measurement claims and the hierarchy of causal evidence should not be conflated. One classifies the inference being made; the other evaluates the evidentiary basis supporting a causal inference. A causal or incremental claim can be conceptually stronger than an attribution claim while still being supported by weak, moderate or strong evidence depending on how the counterfactual was identified.

Measurement providers should state clearly what they have **observed**, what they have **inferred** and what assumptions **connect** the two. Reporting should distinguish impressions from responses, responses from conversions, attributed conversions from incremental conversions, and attribution credit from causal effect.

Advertisers should likewise begin with the decision they are trying to make, rather than the metric a particular platform happens to provide. A useful measurement brief should specify the intervention, population, outcome, time horizon and level of causal confidence required.

The ambition should not always be to produce the most granular possible answer. As Chapter 1 established, granularity and confidence can be different things. A credible estimate that a campaign generated between 8,000 and 12,000 incremental sales may be more decision-useful than a model allocating 10,347.6 conversions precisely across hundreds of individual touchpoints without a credible counterfactual.

As discussed above, a practical assessment can be organized around the same three dimensions:

1. What was measured?

- What creative and advertising treatment was delivered, and did it vary materially across the populations or conditions being compared?
- What exposure, response or outcome was actually observed?
- At what unit was it measured, and over what period?

2. What is being claimed?

- Is the result descriptive, associational or predictive, attributed, or causal/incremental?
- What exactly is being attributed to advertising?
- At what level – campaign, channel, publisher, audience, creative, household or impression – is the claim being made?

3. How strong is the supporting evidence?

- What counterfactual and identification strategy support any causal or incremental interpretation?
- What assumptions, missing data or selection effects could materially affect the result?
- What uncertainty surrounds the estimate, and at what level of granularity does the evidence remain credible?

Journey observability, prior demand, identity fragmentation, decision complexity and time horizon affect both the claims that are useful and the strength and granularity of evidence required to support them. The same measurement approach can support very different levels of inference in different consumer journeys.

Chapter 3:

How consumer journeys shape the attribution challenge

Not all conversions are equally attributable. A click followed seconds later by an online purchase presents a very different measurement problem from a habitual grocery purchase, a new-car decision made over several months or a subscription influenced by years of accumulated brand experience. The same exposure-to-outcome methodology can be highly informative in one journey and deeply misleading in another.

The reason is not simply that some journeys are longer. They differ in prior demand, observability, identity continuity, competitive influence, decision complexity and the time over which advertising can work.

Consumer journeys are useful here not as causal models, but as stress tests for attribution. They show when an observed path is likely to contain meaningful evidence – and when it risks becoming a misleadingly complete-looking account of an incomplete causal process.

What is a consumer journey?

At its simplest, a consumer journey or path to transaction is the sequence of states, influences, interactions and actions that precede a defined consumer outcome. In digital attribution, the term is often used more narrowly to describe the chronological sequence of observable touchpoints preceding a conversion.

A broader definition is needed for outcomes measurement. Consumer journey means the sequence of states, influences, interactions and actions preceding a defined outcome – not only the touchpoints visible to a digital attribution system. A consumer journey may include advertising exposures, searches, website visits, retailer interactions and other events that can be observed in data. But it also includes influences that may never appear in an attribution record: previous experience with the brand, accumulated advertising memories, recommendations, competitor activity, price, distribution, product availability, habitual behavior and the consumer's underlying needs and motivations.

Consumer journeys are useful here as a diagnostic lens, not as a causal model. Outcomes measurement solutions do not necessarily need to reconstruct every exposure and intermediate action preceding a transaction – a more complete path can improve description without establishing why the outcome occurred.

Consumer journeys differ in ways that directly affect attribution. Some begin with high levels of prior demand or purchase intent; some are short and relatively observable, while others unfold over longer periods and across multiple media, devices, retailers and offline environments. In some cases, advertising exposure and subsequent outcomes can be connected relatively clearly; in others, identity gaps and incomplete data leave important parts of the journey unobserved. These differences affect both what can be measured and how confidently an observed outcome can be attributed to advertising. Creative variation can add further complexity because consumers may not all have received the same advertising treatment.

The various dimensions do not move together. A short digital path can be highly observable while remaining causally difficult because advertising is targeted toward consumers who already intend to buy. A habitual purchase can occur almost immediately while reflecting years of accumulated preference. A longer purchase process may leave many observable signals without revealing the incremental contribution of individual contacts. Conversely, a randomized experiment can estimate the effect of a defined intervention without reconstructing the complete path through which that effect occurred.

Path data nevertheless has practical uses. It can describe observable sequences, identify where measurement coverage is lost, distinguish demand-capture activity from earlier influences that warrant further investigation, support campaign diagnostics and generate hypotheses for subsequent tests. Its role should generally be descriptive and hypothesis-generating unless a credible causal design supports a stronger interpretation.

Television illustrates these challenges particularly clearly. TV is widely used to build brands and create or reinforce demand, and its effects may subsequently appear through activities that are much easier to observe digitally: a consumer searches for the brand, visits its website, reads reviews, visits a retailer or eventually makes a purchase. In such cases, the measurable digital activity may be partly a consequence of an earlier television exposure rather than an independent source of demand. A path-based attribution system that begins with the search or website visit can give disproportionate credit to the channel that captures the consumer's response while missing the advertising that helped stimulate it. This interaction between television and later search is one reason that treating channels independently can produce a misleading account of how advertising works.

At the same time, consistently connecting television exposures with those subsequent behaviors is intrinsically difficult. Linear television exposure has historically been measured through audience currencies rather than persistent person-level identifiers, while CTV and streaming introduce their own fragmented device, platform and account identifiers. The television exposure may be recorded against a set-top box, smart TV, streaming account or household; the subsequent search may occur on an individual's phone or laptop; and the transaction may be attached to a retailer account, loyalty card or another household member. Identity resolution can connect some of these records, but a household-level connection is not the same thing as proof that the same individual saw the advertisement, conducted the search and made the purchase.

Television effects can extend farther through the consumer journey than television exposure can be tracked. When television builds memory, familiarity, preference or future demand rather than prompting immediate response, more observable downstream channels can receive credit for effects generated partly upstream. Conversely, the existence of a TV exposure before a search or purchase does not prove that television caused that behavior; some of the demand would have existed anyway. The task is not simply to connect more events, but to distinguish demand creation, demand reinforcement and demand capture – and to estimate the counterfactual.

Intention is part of the attribution problem

Prior demand is a core characteristic of any consumer journey. Consumers can enter a campaign with an explicit purchase plan, learned preferences, habitual behavior, latent needs or other predispositions that make them more likely to respond before the focal advertising occurs. Those predispositions can themselves reflect previous advertising and brand experience, so observed intent may be both a predictor of subsequent behavior and an outcome of earlier marketing influences. The practical implication for attribution is straightforward: measurement should distinguish prior propensity to respond from change caused by the advertising rather than treating consumer intent as an exogenous starting point.

Journey complexity depends on considerably more than the number of recorded touchpoints. Relevant characteristics could include:

- **The consumer's prior state:** whether a need, preference, habit or explicit purchase intention existed before the advertising exposure, and how strongly.
- **The time between advertising exposure and the eventual outcome,** including whether important effects may precede or extend beyond the measurement window.
- **The number of environments through which the consumer moves,** and whether any organization can observe them consistently.
- **The extent to which exposures and outcomes can actually be connected,** including whether the data refer to the same person, device, account or household.

- **The influence of previous brand experience and purchase history**, which may be fundamental to estimating what would have happened without the advertising.
- **The importance of non-advertising influences**, including price, distribution, recommendations, reviews, competitive activity and retailer merchandising.
- **The mechanism through which the advertising is expected to work**: immediate response, creation or acceleration of demand, reinforcement of memory and preference, brand switching, retention or some combination of these.
- **The feasibility of constructing a credible counterfactual**, through randomization, quasi-experimental methods or appropriately validated causal models.

Put another way, **consumer journeys in today's marketplace vary widely and often markedly in their structure, duration, observability and underlying decision processes**. Some are compressed into minutes and take place largely within a single digital environment; others unfold over weeks or months, span multiple media and commercial environments, and involve a mixture of conscious deliberation, habitual behavior and accumulated brand effects. The attribution problem is not a single problem. Its difficulty depends heavily on the type of journey being measured.

The six stylized journey types below illustrate how measurement conditions change the claim that can credibly be made. They are not intended as a taxonomy of journeys that measurement systems should attempt to reconstruct in full.

Exhibit: Six consumer journeys and their measurement characteristics

Consumer context	What pathway data can usefully show	Principal inferential risk	Most useful decision evidence
1. Direct, short, observable response	Immediate sequence from exposure to response	Prior intent and algorithmic selection	Randomized holdout where feasible
2. Short, cross-environment response	Household/device progression across environments	Identity mismatch and incomplete outcome coverage	Household/device experiment or geo test
3. Research-led decision	Search, review and information-gathering patterns	Confusing demand capture with demand creation	Bounded experiments plus broader cross-channel evidence
4. Omnichannel purchase	Observable overlap across platforms and commerce environments	Missing environments, overlapping platform claims and non-comparable identities	Campaign/package experiments, quasi-experiments and MMM
5. Long-consideration purchase	Funnel movement and intermediate response	Long lags, attribution-window sensitivity and external influences	Longer-horizon experiments, MMM and brand/customer evidence
6. Habitual/low-involvement purchase	Customer state, purchase frequency and repeat behavior	High baseline demand and accumulated brand effects	Large-scale RCT/geo evidence plus longitudinal purchase history

1. Direct, short and observable journeys

Advertisement → Click → Immediate purchase

Direct-response journeys provide strong opportunities for granular descriptive attribution because exposure, response and conversion may occur within one environment or be connected through a common account, browser or device identifier. Pathway data can establish the observed sequence with considerable precision and support reporting by campaign, placement, creative or audience.

The principal inferential risk is prior intent and algorithmic selection. A consumer who responds may already have been likely to purchase, and targeting may further concentrate advertising among consumers predicted to convert. Where advertising eligibility can be randomized, a holdout provides stronger evidence about how many additional conversions the intervention generated.

The appropriate division of labor is clear: **pathway evidence describes the visible route to response; experimental evidence tests whether the advertising changed the response.**

2. Short but cross-environment journeys

CTV exposure → Website or app visit → Retailer or marketplace → Purchase

A journey can be short without being consistently observable. A CTV exposure may be recorded at device or household level, subsequent activity against a browser or account, and the transaction by an advertiser or retailer. Identity resolution can connect some of these records and materially improve coverage, but the underlying units and populations should remain explicit.

Pathway evidence is useful for showing which observable events can be connected and where measurement breaks occur. Its principal limitations are identity mismatch and incomplete outcome coverage. Where the decision requires a causal estimate, household- or device-level randomized designs, or geographic experiments where individual randomization is unavailable, can provide a more appropriate comparison.

Journey duration and journey observability are separate dimensions.

3. Research-led journeys

TV → Search → Reviews → Retailer → Purchase

Research-led journeys illustrate the distinction between **demand creation and demand capture**. Search, reviews, retailer activity and other signals close to purchase may be highly observable precisely because the consumer is already moving toward a decision. Their proximity to conversion does not establish that they created the underlying demand.

Pathway data can reveal how consumers gather information and which observable interactions commonly precede purchase. Bounded experiments can provide stronger evidence about particular advertising interventions, while broader cross-channel evidence can help assess activity whose influence developed earlier in the journey.

The measurement objective should be to understand how observable research behavior relates to the decision, without automatically assigning the greatest causal contribution to the touchpoints closest to the transaction.

4. Omnichannel journeys

Media and platforms → Retailers and marketplaces → Online or physical-store purchase

In an omnichannel journey, different platforms, retailers and data partners may each observe legitimate but incomplete parts of the same process. Identity systems may not align, some environments may be missing entirely and several providers may independently claim credit for overlapping outcomes.

Pathway data remain valuable for identifying observable overlap and understanding how consumers move among measurable environments. They are less reliable as a complete causal decomposition of the journey. Campaign- or package-level experiments, credible quasi-experiments and MMM can provide more useful evidence for broader allocation decisions, while platform-level attribution remains valuable for diagnostics and optimization within defined boundaries.

The central risk is not simply missing data; it is treating several partial views as though they were mutually exclusive causal accounts of the whole journey.

5. Long-consideration journeys

TV → CTV → Website visits → Dealer or adviser → Purchase months later

Long-consideration purchases combine extended time horizons with numerous advertising and non-advertising influences. Pricing, product availability, recommendations, competitor activity and repeated media exposure may all change during the period in which the decision develops. The result can also be highly sensitive to the attribution window selected.

Pathway evidence can still reveal intermediate responses and movement through observable stages of consideration, but touchpoint-level causal allocation becomes increasingly difficult as the lag and number of competing influences increase. Longer-horizon experiments where feasible, MMM and brand or customer evidence are more appropriate complements for consequential investment decisions.

In these journeys, a more aggregate causal estimate may be substantially more reliable than a detailed reconstruction of credit across individual contacts.

6. Habitual or low-involvement journeys

Repeated advertising → Routine purchase occasion → Habitual purchase

Habitual purchases demonstrate why a short journey is not necessarily an easy causal problem. The transaction may occur immediately after advertising while being driven largely by established preference, purchasing routines and baseline category demand accumulated over much longer periods.

Pathway data can help identify customer state, repeat purchasing and exposure patterns, but exposure close to a routine transaction may capture demand that already existed. Large-scale randomized or geographic experiments, combined with longitudinal purchase history, can provide stronger evidence about whether current advertising changes purchase probability, frequency, switching, retention or another relevant outcome.

Where sample sizes permit, effects should also be examined by prior customer state. A high observed conversion rate among loyal buyers need not imply a large incremental advertising effect.

Journey complexity changes both the evidence and the claim

Attribution challenges do not arise from one problem alone. Different journeys create different combinations of observational and causal uncertainty.

Exhibit: Different journeys present different attribution challenges

Journey characteristic	What it changes
Longer delay between exposure and outcome	Increases the number of intervening influences and the importance of carryover
More platforms and environments	Reduces the likelihood that any one system observes the complete journey
Different exposure and outcome units	Creates uncertainty about whether the same person received the advertising and produced the outcome
Stronger prior preference or habitual demand	Increases the risk that existing demand is attributed to current advertising
More addressable targeting	Can improve exposure control while also increasing selection bias in observational comparisons
More channel interaction	Makes isolated channel credit harder to interpret
Less frequent or higher-value outcomes	Reduces statistical power and makes experiments more demanding
Greater variation in individual pathways	Makes complete touchpoint-level causal decomposition less feasible

The six categories should not be read as a simple progression from easy-to-measure to difficult-to-measure journeys. Measurement difficulty depends on the particular characteristics of each journey, not simply its length or number of touchpoints. A direct digital journey may be highly observable but heavily affected by prior intent. A CTV-to-purchase journey may be short but depend on household-level identity matching. A habitual purchase may involve few visible steps but reflect years of accumulated brand preference. A long-consideration purchase may be observable at its beginning and end while much of the intervening decision process remains unknown.

The appropriate measurement ambition changes with the journey. For a direct, observable response, it may be reasonable to estimate incremental conversions at campaign or audience level. For a short cross-environment journey, household-level incrementality may be supportable even where person-level attribution is not. In an omnichannel or long-consideration journey, the strongest conclusion may concern the incremental effect of a campaign package or a change in media strategy rather than the contribution of each impression. In habitual categories, the analysis may need to focus on changes in purchase probability, customer acquisition, switching or retention rather than assigning routine purchases to recent exposures.

Implications

Across these six contexts, the underlying principles of causal inference do not change; the usefulness of different forms of evidence does. As pathway observability declines, prior demand becomes more important, identity becomes less continuous or the decision horizon lengthens, credible conclusions may need to become more aggregate, more probabilistic or more dependent on complementary methods.

The objective is not to reconstruct every consumer journey in full. It is to determine what the observed pathway contributes to the decision and where stronger evidence is required.

A credible estimate of campaign, channel or strategy incrementality can be more decision-useful than a precise-looking allocation across hundreds of individual touchpoints.

The practical test is: What decision can the observed pathway evidence support, and what additional causal evidence is required before acting on it?

Chapter 4:

Putting it all together – the measurement confidence gap

The most detailed answer is not always the most credible answer. Modern outcomes systems can report performance by publisher, audience, household, creative, placement and even impression. Yet the evidence establishing that advertising caused the outcome may be reliable only at the campaign, channel or market level.

That mismatch is the measurement-confidence gap: the distance between the precision or certainty implied by a reported result and the level at which the underlying evidence can actually support the claim.

Measurement confidence does not belong to a methodology, dataset or provider in the abstract. It belongs to a specific claim, supported by a particular combination of data, research design, assumptions and evidence. The difficulty of supporting that claim can also vary substantially with the consumer journey, including prior demand, identity fragmentation, time horizon, competing influences and the observability of the path.

The governing discipline is simple: **The strength, scope and granularity of the claim should never exceed the strength, scope and granularity of the evidence.**

The unit of evaluation is the claim

There is no single answer to the question: How good is attribution? The same measurement environment can support several claims with very different evidentiary strength. A well-designed experiment may provide strong evidence that a campaign generated incremental sales without establishing the separate causal contribution of every impression within the campaign. A path-based attribution model may allocate credit across many publishers, audiences and touchpoints without establishing that those allocated shares represent incremental effects.

These outputs are not necessarily in conflict. They answer different questions. The appropriate unit of evaluation is the claim, not the methodology in the abstract. The relevant questions are what outcome is being measured, what relationship to advertising is being asserted, what comparison or counterfactual supports that interpretation, and at what level the evidence remains credible.

This distinction matters because apparently identical metric labels can conceal materially different quantities. Two providers may both report “ROAS.” One might divide revenue observed among exposed households by advertising spend. Another might statistically adjust for observable differences between exposed and unexposed consumers. A third might use a randomized holdout or another credible counterfactual design to estimate incremental revenue.

The numbers may look comparable while answering different questions. The same problem can arise at creative, publisher or audience level. A system may report a detailed difference in performance without possessing equally detailed causal evidence about why that difference occurred. Greater reporting resolution can improve diagnosis and optimization, but it should not silently upgrade the evidentiary status of the result.

A mature outcomes marketplace requires clarity not only about the metric being reported, but about the claim attached to it and the evidence supporting that claim.

Measurement confidence is a decision standard

Measurement confidence matters because measurement is used to make decisions.

Outcomes evidence can influence whether an advertiser increases or reduces investment, changes a media mix, reallocates budget across sellers, changes targeting or creative, extends a campaign, accepts an outcomes guarantee or changes its underlying measurement architecture. The standard of evidence should be proportionate to both the strength of the claim and the consequence of the decision.

A directional association may be entirely adequate for generating a hypothesis, diagnosing campaign performance or informing a relatively low-risk tactical adjustment. A substantial budget reallocation justified on the basis that one channel or seller generated superior **incremental** returns requires stronger evidence.

This does not mean that every decision requires a randomized experiment. Nor does it mean that descriptive or observational evidence is inherently weak or unusable. The relevant question is whether the evidence is strong enough for the interpretation and action attached to it.

This changes the objective of measurement. The goal is not simply to produce the most detailed possible number. It is to produce evidence that decision-makers understand well enough, and trust sufficiently, to use appropriately.

A highly granular report that does not support the decision being made has limited practical value. A less granular estimate can be more useful if its causal basis is stronger, its uncertainty is explicit and its scope is aligned with the decision.

Granularity and confidence are different properties

Modern measurement systems can produce outcomes estimates at increasingly fine levels: publisher, platform, audience, household, creative, placement and individual impression.

That capability is valuable. But **resolution of output and resolution of confidence are not the same thing**.

Suppose a randomized experiment provides credible evidence that an integrated campaign generated incremental sales. The experimental design may identify that effect only at campaign level. A subsequent attribution model may then distribute the estimated effect across channels, audiences, creatives or individual exposures.

The resulting allocation is more granular. It may also be useful. But some of the additional detail has been produced through modeling assumptions rather than independently established by the experiment.

Two levels of resolution need to be distinguished:

Resolution of output is the finest level at which a system can produce a number.

Resolution of confidence is the finest level at which the available evidence remains sufficiently strong to support the interpretation attached to that number.

The two may coincide, but they need not. For some short and highly observable journeys, the evidence may support relatively fine-grained conclusions. For more fragmented or long-consideration journeys, the strongest causal conclusion may sit at campaign, channel, audience or geographic level even where the technology can generate impression-level allocations.

The implication is not that measurement should become less granular. Granular observational evidence can be extremely valuable for diagnostics, prediction and optimization. The requirement is simply to **avoid allowing the precision of the reported number to imply a level of causal knowledge that the evidence has not established**.

Existing industry standards provide an important foundation

The industry is not starting from a blank sheet. In 2022, the Media Rating Council published its Outcomes and Data Quality Standards, developed with support from the ANA, 4As and ACA and a broad cross-industry working group. The standards establish requirements across attribution, multi-touch attribution, test-and-control measurement, marketing-mix modeling, data integration, identity, modeling, disclosure and audit.

The framework does not prescribe a single methodology. Instead, it requires methods to be appropriate to their intended use, empirically supported and sufficiently transparent for users and auditors to understand their assumptions and limitations. Attribution weights and lookback periods should be supported and disclosed; test-and-control designs should establish treatment integrity and credible comparison groups; MMM models should address relevant non-media influences and support the granularity at which results are reported; and data quality, matching, coverage and privacy-related limitations should be assessed and disclosed¹³.

These principles anticipate many of the distinctions developed in this paper. Their existence also illustrates a broader challenge: establishing sound methodological expectations is different from achieving consistent adoption, implementation and interpretation across a fragmented marketplace.

Different methods contribute different forms of evidence

The same principle changes how measurement methodologies should be compared. Attribution, experiments, quasi-experiments, marketing-mix modeling, brand research and customer analytics should not be treated as competing attempts to produce one universal answer. They operate at different levels, depend on different assumptions and illuminate different parts of advertising effectiveness.

Observational attribution can describe pathways, associations and response patterns at substantial scale and granularity. Experiments can provide stronger causal evidence about defined interventions where credible treatment and comparison conditions can be created. Quasi-experimental designs can extend causal inference where randomization is unavailable, subject to stronger assumptions about identification. Marketing-mix modeling can examine media alongside other business drivers across channels and longer periods. Brand and consumer research can measure changes in attention, memory, awareness, consideration and preference. Longitudinal customer and retailer data can reveal how purchasing, retention and customer value develop over time.

The purpose of combining these sources is not to force them into artificial agreement or to average their outputs together. It is to determine where they corroborate one another, where they answer genuinely different questions and where important uncertainty remains.

An experiment showing that a campaign increased sales should not automatically be interpreted as validating every touchpoint-level allocation subsequently produced by an attribution model. Equally, an attribution model identifying stable pathway patterns may generate valuable hypotheses about audiences, publishers or creative even where it cannot independently establish causal contribution.

Where the claim itself is causal, evidence supported by stronger identification should generally receive greater weight. This does not make other forms of evidence less useful. It gives each form of evidence the role its design can credibly support.

The strongest measurement system is not necessarily the one producing the largest number of outputs. It is one that **matches the method to the question, distinguishes observation from inference and makes the evidentiary status of the resulting claim explicit.**

¹³ Source: Media Rating Council (MRC), *MRC Outcomes and Data Quality Standards*, Final, September 2022. Developed with the Association of National Advertisers (ANA), American Association of Advertising Agencies (4As) and Association of Canadian Advertisers (ACA). The standards cover outcomes measurement, attribution and multi-touch attribution, test-and-control measurement, marketing/media mix modeling, modeling and data adjustment, data quality, reporting, disclosure and auditing.

A practical confidence test for outcomes claims

The three dimensions established in Chapter 2 provide a simple structure for evaluating a material outcomes claim.

1. What was measured?

The user should be able to establish:

- **What advertising treatment was delivered.** What counted as exposure, at what unit was it measured and did material differences in creative, frequency or delivery exist across the conditions being compared?
- **What outcome was measured.** Is the result an intermediate response, brand effect, behavioral outcome or economic outcome, and over what period?
- **What measurement boundary applies.** Which consumers, advertising exposures, transactions, retailers, platforms or other relevant activity are included or excluded, and how were exposure and outcome records connected?

2. What is being claimed?

The result should make clear:

- **Whether the claim is observational, associational or predictive, attributed, or causal/incremental.**
- **What advertising intervention or activity the claim concerns.** Is the conclusion about a campaign, channel, publisher, audience, creative strategy or individual exposure?
- **At what level the claim is being made.** A campaign-level causal estimate should not automatically become an impression-level causal claim because a model can allocate the result at that resolution.

3. How strong is the evidence?

For claims carrying a causal or incremental interpretation, the user should understand:

- **What supports the counterfactual:** What randomized, quasi-experimental or other identification strategy distinguishes advertising effects from prior demand and other influences?
- **What material assumptions and uncertainties remain:** These may include selection, identity error, incomplete exposure or outcome coverage, attribution-window choices, model specification, statistical uncertainty and generalization beyond the population or period studied.
- **Whether the conclusion has been tested against stronger or independent evidence where appropriate:** Experimental evidence, alternative model specifications, sensitivity tests or other external validation can materially affect the confidence warranted by the result.

These questions do not prescribe a single methodology. They provide a common discipline for deciding **what a result means, how strongly it should be stated and whether it is sufficient for the decision being made.**

Measurement confidence can become a marketplace capability

The measurement-confidence gap is not only a methodological problem. It has commercial implications for television and video.

Integrated digital platforms can possess an important structural advantage because media delivery, identity, first-party signals, conversion data, measurement and optimization can often operate within closely connected environments. Fewer organizational and technical boundaries can make evidence easier to connect and use and can support faster activation and experimentation.

That advantage remains bounded by the measurement environment. A platform may have exceptionally rich exposure, behavioral and transaction data within its own systems while observing less of what happens elsewhere, including activity across other media, retailers or offline environments. Integration also does not automatically produce less biased attribution, stronger causal identification or more representative evidence.

The TV and video ecosystem is organized differently, across multiple sellers, platforms, identity systems and measurement providers. Its opportunity is to combine rich outcome signals with clearer evidence about what those signals establish, stronger causal testing where appropriate and greater consistency in how credible outcomes measurement can operate across participating systems.

This is not a presumption that stronger causal evidence will favor television. A credible experiment can identify positive, negligible or negative effects. Its value lies in increasing confidence that subsequent investment decisions are responding to advertising effects rather than to measurement artifacts or pre-existing demand.

Addressable and CTV environments can create useful opportunities in this respect because advertising eligibility can sometimes be varied at household, device or geographic level. But the same discipline applies: strong experimental evidence about a defined CTV intervention does not automatically create a complete causal account of every exposure within the campaign or of the wider television plan.

The more general opportunity is to make evidentiary confidence itself part of the value proposition of an accountable outcomes marketplace. Advertisers do not simply need more outcome signals. They need to know which signals are sufficiently credible for which decisions.

From measurement confidence to marketplace assessment

The resulting discipline is simple: the language, scope and granularity of an outcomes claim should not outrun the evidence supporting it.

The relevant question is not whether a methodology or provider is intrinsically “good” or “bad”, but what a specific combination of advertising activity, exposure data, outcome, population and research design has actually established – at what level, with what uncertainty, and with sufficient confidence to support which decision.

That provides the basis for assessing the contemporary U.S. attribution and outcomes ecosystem, where increasingly rich exposure, identity and commercial signals coexist with substantial variation in measurement boundaries, methodology, causal evidence and disclosure.

Chapter 5:

The U.S. marketplace today – abundant signals, uneven proof

The U.S. market does not have a shortage of outcomes data. It has a consistency problem in what those data are understood to prove.

Across linear television, addressable TV, streaming and CTV, advertisers can increasingly connect exposure with website activity, search, store visits, subscriptions, purchases and other commercial behavior. Audience-measurement companies, outcomes specialists, publishers, platforms, identity providers, clean rooms, retailers, DSPs, ad servers and modeling providers now contribute different pieces of an increasingly connected measurement stack.

Taken together, these developments amount to a sophisticated and increasingly operational outcomes-measurement ecosystem. They represent genuine advances in what advertisers can observe, connect, test and use in media decision-making. The remaining challenges concern the interpretation, evidentiary strength and cross-market usability of the capabilities that now exist. There are still important distinctions between exposure and attention, identity and causality, attribution and incrementality, and granular output and causal confidence.

The relevant question for each new capability is not simply *what can it measure?* It is *what does that measurement allow the buyer to conclude?*

Five structural developments are reshaping the U.S. outcomes-measurement marketplace. Each changes what the market can observe or do, while leaving important evidentiary questions unresolved. The developments can be assessed against the same three-part test: what capability has improved, what measurement value that improvement creates, and what inference still remains unsupported without additional evidence.

Exhibit: What has changed – and what has not

Structural development	Capability gained	Measurement value created	What it still does not establish
1. Audience and outcomes measurement are converging	Exposure, reach, identity and commercial outcomes increasingly sit within connected analytical systems.	Media delivery and subsequent response can be evaluated together, improving diagnosis of reach, frequency, audience composition and commercial outcomes.	Integrated data do not by themselves establish whether advertising caused the observed outcome.
2. Publishers and platforms increasingly measure their own media	First-party exposure, identity, behavioral and commercial signals can operate inside the same environment.	Richer closed-loop measurement, faster optimization and stronger opportunities for experimentation within defined platform boundaries.	Measurement depth within one environment does not establish campaign-wide effects or independent validity across the wider media system.
3. Clean rooms and identity systems are becoming infrastructure	Advertiser, publisher and retailer datasets can increasingly contribute to common analyses without unrestricted exchange of row-level records.	Better governed data collaboration, cross-party linkage, experimentation, reach/frequency analysis and MMM inputs.	A successful data join does not create a counterfactual or eliminate identity, coverage and selection problems.
4. Attribution, experiments and MMM are becoming more integrated	Experimental and aggregate evidence can be used to test, calibrate or constrain scalable attribution models.	Greater potential to combine causal anchors with scale, granularity and cross-media context.	Granular model outputs do not inherit causal status beyond the evidence used to calibrate or validate them.
5. Measurement is moving into execution	Outcome signals increasingly influence bidding, targeting, creative, inventory and budget decisions while campaigns are running.	Faster learning and more responsive media allocation.	Optimization against an observed or predicted outcome does not prove optimization for incremental business effect.

Existing standards provide an important foundation but adoption remains uneven

The industry is not starting from a blank sheet. In 2022, the Media Rating Council published its *Outcomes and Data Quality Standards*, developed with support from the ANA, 4As and ACA and a broad cross-industry working group. The standards establish requirements across attribution and multi-touch attribution, test-and-control measurement, marketing-mix modeling, model governance, data integration and quality, reporting, disclosure and audit¹⁴.

The framework does not prescribe a single methodology. Instead, it establishes disciplines around fitness for purpose, empirical support, validation, transparency and auditability. Attribution weights and lookback periods should be supported and disclosed; experimental designs should address treatment integrity and the comparability of treatment and control groups; MMM should account appropriately for relevant non-media influences and support the granularity at which results are reported; and material limitations in data coverage, matching and methodology should be made visible.

These standards anticipate many of the evidentiary disciplines described in this paper. But the existence of a standard is different from consistent marketplace adoption. Formal MRC accreditation for outcomes measurement remains limited. MRC's July 2026 accreditation-status report for major social and retail-media platforms shows none of the platforms covered by the report accredited or in process for its Outcomes category¹⁵. The wider MRC digital-accreditation directory identifies some accredited outcomes-related services but does not indicate broad market-wide outcomes accreditation.

The distinction is important. Much of the methodological foundation for credible outcomes measurement already exists. The remaining challenge is to understand how consistently existing principles are being implemented and disclosed, where material gaps remain, and what would make stronger practices sufficiently useful to become embedded in routine advertiser, agency, publisher and measurement workflows.

1. The boundaries between audience measurement and outcomes measurement are eroding

One of the most important structural developments is the erosion of the traditional boundary between measuring who received advertising and measuring what happened afterwards. Exposure, reach, frequency, identity and commercial outcome data increasingly operate within connected analytical environments rather than as separate measurement products.

This convergence matters because exposure measurement is increasingly becoming an input into effectiveness analysis rather than a separate end product. Delivery, audience, response and commercial outcomes can be connected within the same analytical environment, making it easier to examine who was reached, what happened subsequently and how those outcomes varied across audiences, publishers, campaigns or creative treatments.

What has improved is the ability to connect media delivery with subsequent behavior within the same analytical system.

That can materially reduce the distance between audience measurement and outcomes analysis. It can also improve diagnosis. An advertiser may be able to examine whether differences in sales performance coincide with differences in reach, frequency, audience composition or exposure distribution rather than treating delivery and outcomes as unrelated datasets.

14 Media Rating Council (MRC), *MRC Outcomes and Data Quality Standards*, Final, September 2022. Developed with the Association of National Advertisers (ANA), American Association of Advertising Agencies (4As) and Association of Canadian Advertisers (ACA). The standards address outcomes measurement, attribution and multi-touch attribution, test-and-control measurement, marketing/media mix modeling, modeling and data adjustment, data quality, reporting, disclosure and auditing.

15 Media Rating Council (MRC), *Digital Platforms Accreditation Status*, July 9, 2026. The MRC chart reports accreditation status across selected priority measurement areas for major social and retail-media platforms; the Outcomes category is shown as "not accredited or in process" for the platforms included in the report.

The gain is integration, not automatic causal identification. Bringing delivery and outcome evidence together can improve diagnosis, reveal mismatches between reach and response, and make experimentation easier to execute. But an observed purchase, an attributed purchase and an experimentally estimated incremental purchase remain different forms of evidence. The analytical value of convergence lies in making more of the measurement system visible in one place; it does not, by itself, upgrade the causal status of the claims produced within that system.

2. Publishers and platforms have become measurement companies

A second structural change is the growth of publisher- and platform-controlled measurement.

Large digital and streaming platforms can possess assets that traditional television companies rarely held together: authenticated or partially authenticated audiences, impression-level delivery records, direct technological relationships with users and, in some cases, extensive information about subsequent consumer or commercial behavior.

Amazon illustrates the model particularly clearly. Its advertising environment combines media exposure with commerce signals, while Amazon Marketing Cloud enables advertisers to analyze Amazon advertising information alongside their own data within a controlled environment. Its developing multi-touch attribution methodology goes further by using randomized experiments to help constrain machine-learning estimates of advertising contribution¹⁶.

Amazon is illustrative of a broader platform model in which media delivery, identity, first-party or commerce signals, measurement, experimentation and optimization can operate within closely connected technical environments. The strategic significance lies less in any single provider than in the structural advantage created when these components are integrated.

The principal advantage is **measurement depth within the platform environment**. A platform may be able to observe its own advertising delivery at high resolution, connect exposure with authenticated or first-party signals, measure subsequent activity and, where advertising eligibility can be controlled, support randomized experimentation. In some environments, targeting, delivery, outcome measurement, experimentation and optimization can occur within the same technical system.

That creates measurement opportunities that traditional television rarely possessed. Two important limitations nevertheless remain:

1. The first is **scope**. Even an exceptionally rich platform dataset normally represents only part of the consumer's media and purchase journey. An advertiser may obtain a detailed view of activity within one platform while retaining limited evidence about advertising exposures, searches, retailer interactions, competitive activity or purchases occurring elsewhere. Greater depth within one environment does not produce a complete view across environments.
2. The second is **governance and independence**. In some cases, the organization selling the advertising may also define exposure, operate or select the identity system, control access to the underlying data, determine the available outcome and methodology and report the result.

Platform environments can provide unusually deep evidence inside a defined measurement boundary while still offering an incomplete view of the advertiser's total campaign. That distinction should be explicit.

A platform may possess rich first-party signals, highly precise exposure data and strong experimental capability without thereby establishing the effect of activity occurring outside its environment. Similarly, the quality of the evidence and the independence of the measurement are separate questions. Closed-loop capability is a structural advantage; it is not, by itself, a guarantee of campaign-wide causal validity.

¹⁶ Source: Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero and German Schnaidt, *Amazon Ads Multi-Touch Attribution* (August 2025).

The distinction is important. Integration can improve data continuity, observability, experimental feasibility and speed of activation. Whether it improves the quality of a particular effectiveness estimate is a separate empirical question. A closed environment can still suffer from selection bias, incomplete outcome coverage, inappropriate comparison groups or causal claims that exceed the intervention actually tested.

3. Clean rooms improve access to evidence, not the logic of causation

Data clean rooms are increasingly part of the infrastructure connecting advertiser, publisher and retailer evidence. They can allow datasets controlled by different organizations to contribute to common analysis without unrestricted exchange of row-level records. In advertising measurement, this can support exposure-to-conversion analysis, cross-publisher reach and frequency, advertiser and publisher data collaboration, aggregated inputs for MMM and, where the design permits, incrementality testing.

The capability that has improved is governed access to evidence that was previously separated. This can materially increase measurement coverage and make analyses possible that would otherwise require direct transfer of sensitive data or could not be conducted at all.

What a clean room does not provide is a counterfactual. Connecting an exposure record with a purchase record can improve confidence that the two records belong in the same analysis; it does not establish what would have happened without the advertising. A clean room can support causal measurement where treatment assignment, comparison groups, outcome data and research design provide credible identification, but the causal evidence comes from that design rather than from the clean-room environment itself.

The same distinction applies to identity. Better matching can improve linkage confidence while leaving causal confidence unchanged. Match rates, the unit at which records are connected, excluded consumers and missing exposures or transactions remain relevant to interpretation.

Clean rooms should be understood as measurement infrastructure: they can improve access, governance and analytical capability, while the evidentiary status of the resulting claim remains determined by the data, research design, assumptions and counterfactual applied within or around that infrastructure.

4. Attribution, experiments and MMM are beginning to operate as a connected evidence system

An important emerging development is the use of experimental and aggregate evidence to test, calibrate or constrain scalable attribution models rather than treating each methodology as a wholly separate source of truth. This architecture is not yet a uniform market practice, but it is visible in deployed systems and recent methodological research¹⁷.

Experiments are beginning to discipline scalable attribution models

One important development is the use of experimental evidence to test, calibrate or constrain scalable attribution models. Observational models can operate at high resolution and predict conversion effectively while remaining biased about advertising effects. Randomized evidence can provide a stronger causal benchmark, although typically for more bounded interventions.

Amazon's published multi-touch attribution methodology provides a prominent example: randomized experiments are used alongside machine learning and observed shopping signals to inform attribution credit. The significance is architectural rather than provider-specific. Attribution need not infer causation entirely from the same observational relationships it is trying to explain.

¹⁷ **Sources:** Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero and German Schnaidt, *Amazon Ads Multi-Touch Attribution* (August 2025); Brett R. Gordon, Robert Moakler and Florian Zettelmeyer, *Predicted Incrementality by Experimentation (PIE) for Ad Measurement*, NBER Working Paper 35044 (April 2026).

The qualification is equally important. Experimental calibration does not transform every granular model output into an experimentally identified causal effect. An experiment may establish a campaign-level incrementality estimate while the subsequent allocation of that effect across publishers, audiences, creatives or individual touchpoints remains model-dependent.

The capability gained is better-constrained attribution, not causal certainty at arbitrary levels of granularity.

MMM is becoming the cross-media decision layer

The renewed importance of marketing mix modeling reflects the same fragmentation reshaping attribution.

No platform or identity architecture observes the complete movement of consumers across television, streaming, social media, search, retail media, e-commerce and physical stores. Individual-level attribution attempts to reconstruct observable consumer pathways. MMM approaches the problem at a different level, using aggregate variation in media activity, business outcomes and other demand drivers.

Its principal advantage is **breadth**. MMM can place television and video alongside other digital and offline media, pricing, promotions, distribution, seasonality, competitor activity and economic conditions. It can also represent carryover, saturation and longer-term effects that may extend beyond the relatively short windows commonly used in impression-level attribution.

This makes MMM particularly useful as a broader check on conclusions generated within more bounded identity and platform environments. If an attribution system suggests very strong returns from a channel but business-level outcomes show little corresponding variation when investment changes, that discrepancy is analytically important. Conversely, an aggregate model may identify effects from television or other upper-funnel media that are poorly represented in path-based data because those effects develop over longer periods or are subsequently captured by lower-funnel channels.

MMM is becoming more important for a complementary reason: no platform or identity system observes the complete movement of consumers across television, streaming, social media, search, retail media, e-commerce and physical stores. MMM approaches the problem at a broader level, using aggregate variation in media activity, business outcomes and other demand drivers.

Its value is breadth. It can place television and video alongside other media, pricing, promotions, distribution, seasonality, competitor activity and economic conditions, and can represent effects that extend beyond short attribution windows. Experimental findings and more granular exposure data can inform or constrain these models, while MMM can provide wider context for results generated within individual platforms or campaigns.

The remaining limitation is identification. Greater model sophistication does not guarantee that historical data contain sufficient independent variation to estimate every channel effect reliably. MMM contributes a broader cross-media and business-level decision layer; it does not eliminate the need for experimental evidence or become a substitute for granular attribution.

Its emerging role is different: **a broad cross-media and business-level decision layer that can place granular attribution and bounded experimental results within a wider system of demand**. The important development is not that these methods always agree but that they can be used to test, calibrate and constrain one another.

This represents a structural change in the measurement market. Attribution, experimentation and MMM are becoming components of a connected evidence system rather than isolated products competing to produce one definitive answer. This can strengthen triangulation, but integration can also obscure where experimentally identified evidence ends and model-dependent inference begins.

5. Measurement is moving from reporting into media execution

The final structural change is operational rather than purely methodological. Outcomes measurement increasingly influences campaigns while they are running. Publishers, DSPs and other media systems can feed response and outcome signals into audience selection, bidding, frequency, creative selection, inventory allocation and budget optimization. MMM-derived response curves and allocation recommendations can influence planning and subsequent spending decisions as well.

The significance of this shift is that measurement is no longer merely descriptive. Once an outcome signal affects subsequent targeting, bidding or allocation, the properties of that signal become properties of the media-delivery system itself.

The underlying learning loop is increasingly:

Media delivery → measured response → optimization → future media delivery

This can be extremely valuable. Measurement no longer needs to remain a retrospective report produced after a campaign has ended. Signals can be used to improve subsequent decisions while advertising remains active.

The closer measurement moves to execution, the more consequential measurement error becomes. If an attribution system over-credits high-propensity consumers and that signal is fed directly into optimization, the system can allocate more spending toward consumers who were already likely to buy, observe their high conversion rate and reinforce the same bias.

The measurement process can become endogenous to the media-delivery process. Measurement error is no longer confined to the report. It can change the next media decision automatically.

Optimization for observed or predicted conversion is different from optimization for incremental response. A system may improve conversion rate or cost per conversion by concentrating investment among consumers already likely to buy. That can be operationally valuable without demonstrating that the advertising is creating more incremental business value.

The capability gained is speed: measurement can improve media decisions while the campaign is running. The risk is recursive bias: if a measurement system over-credits high-propensity consumers and that signal feeds future delivery, the system can reinforce the same error automatically.

As measurement becomes embedded in execution, causal calibration becomes more consequential rather than less. Advertisers need evidence that improvements in the signal being optimized correspond to improvements in the business outcomes advertising actually changes.

This is one reason experimental calibration becomes important. Where optimization systems learn from attributed or predicted outcomes, advertisers need evidence that improvements in the optimized signal correspond to improvements in the business outcomes advertising actually causes.

Where does this leave the marketplace?

The contemporary U.S. outcomes marketplace is not short of signals. Its central challenge is that improvements in observability, linkage, modeling and automation have advanced unevenly relative to the evidentiary strength and comparability of the claims produced from them.

Across the five developments examined in this chapter, the pattern is consistent. Audience and outcomes convergence provides richer connected evidence, but integration does not establish causation. Platform measurement creates exceptional depth inside defined environments, but not necessarily campaign-wide scope or independent verification. Clean rooms improve governed access to data, but do not create the counterfactual. Experiment-informed attribution can strengthen scalable models, but granular model outputs remain bounded by the evidence used to calibrate them. MMM can provide broader cross-media context but remains dependent on the variation and assumptions available to identify effects. And when outcome signals move directly into optimization, measurement error can become embedded in subsequent media delivery.

The important distinction is between three separate capabilities:

1. **Observability:** how much relevant advertising and outcome activity can be recorded and connected.
2. **Inference:** what the available research design allows the market to conclude about advertising's contribution to those outcomes.
3. **Operationalization:** how quickly and efficiently those signals can be used to influence planning, buying and optimization.

Progress in one dimension does not automatically produce progress in the others. A platform can be highly observable and operationally sophisticated while supporting only bounded causal claims. A randomized experiment can provide strong causal evidence while offering limited granularity or immediate optimization capability. A clean room can improve access to evidence without strengthening the identification strategy applied to it.

This tension is particularly acute in addressable television and CTV, where measurement depth can be exceptionally high within individual environments while identities, data access, exposure definitions and causal evidence remain fragmented across the wider campaign. More measurement within each environment can coexist with less coherence across the market.

Chapter 6:

Addressable TV and CTV: a measurement revolution – or a fragmentation problem?

Addressability solves one of television’s oldest measurement problems and exposes several new ones.

It allows advertisers to select households more precisely, record delivery at far greater granularity, connect exposures with first-party and commercial data and, in some environments, vary advertising experimentally.

That makes addressable TV and CTV unusually powerful environments for outcomes measurement, but it does **not** make them causally self-validating.

Richer identity and impression data improve observability. They do not automatically remove selection bias, create a counterfactual, reconcile identities across platforms or establish the incremental contribution of every impression in a campaign. At the same time, the fragmentation of platforms, identifiers, definitions and data access can make evidence exceptionally deep inside individual environments while leaving the total campaign difficult to interpret. This is in large part a consequence of the marketplace’s successful proliferation of capabilities: many sophisticated systems now exist, but they do not yet operate as a consistently connected whole.

That is the addressable paradox: more measurement inside each environment can coexist with less coherence across the market.

Exhibit: The four dimensions of the measurement confidence gap

Dimension	Measurement question	Principal implication
Addressability and CTV	Can richer, more granular data solve the problem?	Better observability and targeting improve measurement, but do not automatically create causal confidence or cross-market coverage.
Causal evidence and incrementality	Can stronger research designs establish what advertising changed?	Experiments and credible quasi-experiments provide stronger causal anchors, but their conclusions remain bounded by the intervention tested.
Brand and long-term effects	Does stronger short-term causality capture the full value of advertising?	Important effects may emerge through memory, retention, future demand and customer value beyond the next transaction.
Measurement intelligence	How should these different forms of evidence work together?	A mature system requires a connected portfolio of evidence in which methods test, constrain and complement one another.

What addressability genuinely improves

Addressable television changes the unit at which media decisions can be made. Traditional television advertising typically buys audiences probabilistically: advertisers select programs, networks, dayparts or audience packages expected to contain the people they want to reach. Addressability allows an advertiser instead to make different advertising decisions for different households or devices.

Depending on the environment, those decisions can draw upon subscriber information, geographic location, household characteristics, first-party customer data, purchase histories, viewing behavior and other attributes. Deterministic subscriber relationships can provide a particularly useful household-level anchor, allowing advertisers to target known households rather than inferred clusters.

This creates several potentially important benefits:

1. Addressability can **reduce obvious delivery waste**. An automotive advertiser need not advertise a new-car offer to households that recently purchased the vehicle; an existing-customer campaign can be directed towards customers rather than prospects; geographic advertising can be confined to markets in which the product is available.
2. Addressability can improve **reach and frequency management**. Advertisers can potentially suppress households already heavily exposed through other inventory and direct impressions towards households that conventional television has failed to reach. This is particularly valuable where linear television audiences are concentrated among heavier viewers.
3. Addressability can facilitate **creative differentiation**. Different households can receive different products, offers or messages, potentially making the advertising more relevant.
4. Addressability can improve the measurement of **delivery**. Household-level impression records can provide considerably more granular evidence about which households were eligible for, served, or exposed to advertising than traditional program-level audience measurement.

These are genuine advantages of addressability. They concern **who can be reached, how advertising can be delivered and how accurately that delivery can be observed**.

They do not, by themselves, answer a different question: **does the addressable strategy generate greater incremental advertising value than the best realistic non-addressable alternative?**

That distinction matters particularly when addressable selection uses previous purchasing or predictive data. A high post-exposure conversion rate among a carefully selected audience can reflect both advertising response and the audience's pre-existing probability of buying. The richer the targeting data become, the more important it becomes to distinguish the ability to predict purchasing from evidence that the advertising changed purchasing.

Better targeting is not the same as greater incrementality

Addressability can improve audience selection without making observed conversion a causal measure of advertising impact. When targeting uses purchase history or predictive signals, the households most likely to receive advertising may also be the households most likely to buy without it. Observed conversion combines two quantities:

Baseline propensity to buy + incremental advertising effect.

Purchase history is especially useful because it makes that distinction visible. Consider three households:

Household	Previous behavior	Probability of buying without advertising	Probability after advertising	Incremental effect
A	Heavy brand buyer	70%	73%	+3 pts.
B	Occasional category buyer	20%	28%	+8 pts.
C	Non-buyer	3%	6%	+3 pts.

A conventional outcomes report might favor Household A because it converts at the highest rate. Yet Household B shows the largest incremental response. The example illustrates why purchase-history data can be simultaneously valuable for targeting and essential to the counterfactual: the characteristics that make a household attractive to target can also make it more likely to produce the outcome without advertising.

For addressable television, the practical implication is clear: **the richer the targeting data become, the more important it becomes to distinguish predictive targeting from causal measurement.**

What would constitute convincing evidence that addressability works better?

The strongest test is not: **Did targeted households buy more than untargeted households?**

Nor is it: **Did addressable impressions generate a higher attributed ROAS than non-addressable television impressions?**

Both comparisons can be materially affected by audience selection. Addressable advertising may deliberately reach households with different underlying purchase propensities from those reached through broader television delivery.

The stronger question is decision-based and counterfactual: **For the same business objective, eligible population, advertising budget and campaign period, what additional outcomes were caused by the addressable strategy relative to the best realistic non-addressable strategy the advertiser would otherwise have used?**

Answering that question requires an experiment designed around the actual strategic choice facing the advertiser. Wherever feasible, the addressable strategy and its realistic non-addressable alternative should be compared under common campaign conditions, with random assignment determining which strategy is applied.

Ideally, the design would contain three randomized conditions:

1. **Addressable strategy:** the advertiser deploys the addressable approach being evaluated, including the relevant targeting, suppression, frequency-management or creative rules.
2. **Non-addressable comparator:** the advertiser deploys the best realistic conventional television strategy it would otherwise have used, under a comparable budget, campaign period, business objective and creative strategy wherever practical.
3. **Holdout:** the focal television intervention is withheld, while other material campaign activity is kept as comparable as possible across the experimental conditions.

The unit of randomization should reflect the way the media can actually be controlled. Where both strategies can be assigned credibly at household level, household randomization may be appropriate. Where conventional television delivery cannot be controlled household by household, the comparison may instead require randomized geographic or other appropriately clustered units.

The three comparisons answer different questions. Comparing the addressable condition with the holdout estimates the incremental effect of the **addressable strategy as implemented**. Comparing the non-addressable condition with the holdout estimates the incremental effect of the **alternative television strategy**. The randomized comparison between the two active treatments estimates the incremental advantage – or disadvantage – of choosing the addressable strategy rather than the realistic non-addressable alternative under the conditions tested.

The addressable treatment may change several elements at once: audience selection, suppression, frequency, inventory and creative. A comparison between the complete addressable strategy and conventional television estimates the causal effect of choosing one strategy rather than the other; it does not automatically identify which component produced the difference. If the decision requires a component-level answer, the relevant component must itself be varied experimentally.

Creative should be handled on the same principle. If the objective is to compare addressable with conventional delivery, creative should be held sufficiently comparable across conditions wherever practical. If tailored creative is itself part of the decision, creative assignment should become part of the randomized treatment.

The evidence required depends on the claim. Evidence that an addressable campaign generated incremental sales relative to a holdout supports a claim about the effectiveness of that campaign. Evidence that an addressable treatment outperformed a randomized non-addressable comparator supports the stronger claim that the addressable strategy produced an incremental advantage over the alternative under the conditions tested.

Even that conclusion should remain bounded. It applies to the targeting rules, budget, creative, eligible population, campaign period, outcome and non-addressable alternative actually tested. It should not automatically be generalized into a claim that addressable television produces superior incremental returns across advertisers, categories, audiences or campaign designs.

CTV creates extraordinary measurement depth but uneven breadth

CTV extends many of the capabilities created by addressability. Streaming environments can generate impression-level delivery records at enormous scale; logged-in services can possess rich first-party information about subscribers and viewing behavior; and platforms may know which account or device received an advertisement, when it was delivered and what content surrounded it. Exposure can sometimes be connected with advertiser, retailer or other outcome data, while campaign delivery systems can support experimental holdouts.

The genuine advance is depth within the measured environment. Advertisers can observe delivery and frequency more directly, analyze campaigns at household or device level, connect first-party data within controlled environments and, in some cases, estimate incremental effects experimentally.

The critical qualification is **within the measured environment**. The same household may also encounter the campaign through broadcast or cable television, another streaming service, YouTube, social media or online video. Some of those exposures may be visible to the measurement system; others may remain inside another platform or identity environment.

CTV creates **measurement environments of extraordinary richness without creating a universal view of television exposure**. The analytical challenge is no longer simply generating more data within each environment, but reconciling evidence across environments that use different identities, metadata, measurement methods and commercial rules.

Fragmentation occurs across the CTV attribution stack

CTV fragmentation is not simply a question of different identity systems. Different environments can define the exposure differently, observe different populations, connect different outcome universes, permit different forms of measurement and report results at different levels of granularity. Consequently, two attribution systems apparently measuring the same campaign may not actually be measuring the same exposure, population, outcome universe or attribution question.

Exhibit: Fragmentation across the CTV ecosystem

Layer	What can differ across CTV environments	Why it matters
Exposure	Impression qualification, playback conditions, completion, ad format, log availability.	The denominator being attributed may not be equivalent.
Identity	Device, household, account, login or graph-based identity; deterministic and probabilistic matching.	Populations and match coverage can differ.
Outcomes	Advertiser, retailer, transaction, visitation or platform data; impression-, cohort- or modeled-level outputs.	Platforms may observe different outcome universes.
Attribution	Attribution rule/model, measurement window, treatment of multiple exposures and permitted measurement partners.	Similar metrics can represent different analytical claims.
Data access	Raw logs, clean-room access, aggregated outputs or platform-native reporting.	Independent replication and cross-platform reconciliation vary.
Optimization	Post-campaign reporting versus signals available during campaign delivery.	Attribution may play very different roles in decision-making.

Identity, scope and governance remain binding constraints

CTV does not provide a universal identity layer. Services may rely on subscriber accounts, hashed identifiers, device IDs, smart-TV identifiers, IP-based signals or proprietary identity graphs, using deterministic or probabilistic links. The resulting unit may be a device, household or account rather than an individual viewer, and technical linkage does not necessarily mean that records can be shared or reconciled across commercial environments.

Platform boundaries introduce a separate constraint. A publisher may possess highly accurate information about its own advertising delivery without making all of the underlying data available for independent cross-platform analysis. Advertisers can receive sophisticated platform-generated measurement while having more limited ability to reproduce the analysis or reconcile it with evidence from elsewhere.

Clean rooms can create controlled intersections among some of these datasets, but they cannot create an exposure, identity or transaction record that the participating parties do not possess.

Platform attribution creates a related aggregation problem. Separate platforms may each report outcomes connected with their advertising, but those results cannot simply be summed into a total campaign effect. Consumers can appear in more than one system, platforms may observe different outcome universes and multiple advertising exposures may have contributed to the same transaction.

The relevant distinction is between high-quality evidence within a defined measurement boundary and a complete account of the advertiser's television and video campaign.

Measurement coverage is not the same as campaign coverage

The richness of CTV data also creates a representativeness problem. Streaming adoption, platform usage and linear viewing vary across demographic and behavioral groups. The households observed particularly well within a CTV environment may differ from the total television audience and from the population relevant to a particular advertiser.

A campaign aimed at consumers who remain relatively heavy linear-TV viewers, for example, could be measured with exceptional precision among its CTV-exposed households while being observed much less completely among other important parts of the target population. Measurement quality within the observable subset should not be mistaken for representativeness of the whole campaign.

This also changes the appropriate denominator. The relevant question is not simply how much television viewing occurs on internet-connected television sets. It is what proportion of the advertiser's relevant advertising exposures and outcome journey can actually be observed and reconciled. A connected screen can carry linear, virtual MVPD, FAST and streaming content; connectivity alone does not mean that every advertisement shown on the screen generates an interoperable digital exposure record.

CTV has made substantial parts of television dramatically more measurable without yet creating a complete and unified television measurement system.

The transition can simultaneously increase the quantity and precision of available data while making reconciliation across the total television environment more complex. That is not a reversal of measurement progress. It means that progress should be assessed by whether the additional evidence improves decisions about advertising effectiveness – not simply by the volume of impressions, identities or outcome matches available.

These differences create a problem of comparability as well as coverage. Two CTV environments can each produce internally coherent attribution results while using different exposure definitions, populations, identity methods, outcome sources, attribution windows or models. The resulting estimates cannot necessarily be compared simply because they carry the same metric label. Cross-platform reconciliation requires understanding not only the number reported by each system, but whether the systems are measuring sufficiently comparable quantities for the intended decision.

Experiments strengthen causal inference but remain bounded

CTV and addressable environments can create valuable opportunities for randomized holdouts where advertising eligibility can be controlled. A well-designed platform experiment can estimate the incremental effect of a defined intervention for the eligible population, outcome and period tested.

That result should not automatically be interpreted as the effect of the advertiser's entire television campaign or as the effect of moving budget from another channel. Treatment and holdout groups may both receive television or video advertising elsewhere, so the experiment may estimate the incremental contribution of one platform within the surrounding media environment rather than the standalone effect of CTV.

This can be exactly the relevant estimand. The requirement is to describe it correctly and distinguish strong causal evidence within a defined measurement boundary from a claim about the total campaign.

Implications

Addressability changes the measurement frontier without eliminating the underlying measurement problem. It can improve audience selection, treatment control, exposure records, identity linkage and opportunities for experimentation. Within a defined platform or addressable environment, those advances can produce substantially richer and more useful evidence than was historically available in television.

But greater observability is not the same as greater causal confidence, and greater measurement depth within one environment is not the same as complete campaign coverage. Addressable targeting can intensify selection effects; household identity does not establish person-level exposure; platform-level experiments remain bounded by the intervention and population tested; and highly detailed evidence within individual environments remains difficult to reconcile across the wider television campaign.

The lesson is not that greater granularity has failed. It is that granularity solves only part of the measurement-confidence problem.

The next question is whether stronger research designs can provide the causal confidence that richer observational data alone cannot. Randomized and credible quasi-experimental designs offer a different route: rather than trying to infer causality from increasingly detailed observational data, they create or exploit variation capable of supporting a counterfactual.

Chapter 7:

From attribution to incrementality – building credible causal evidence

If observational data cannot reliably answer the counterfactual question, the market needs evidence created by research design rather than inferred from correlation.

Randomization provides the clearest route where it can be validly implemented. It does not answer every question, and it does not need to. Its value is narrower and more important: it can establish whether a defined change in advertising caused a defined outcome to change.

The objective is not to replace attribution with experimentation. It is to create a measurement system in which stronger causal evidence provides anchors against which more scalable but assumption-dependent models can be tested, constrained and updated.

A hierarchy of evidence for causal claims

Different levels of measurement claim need to be distinguished from the strength of the evidence supporting them. Observing an outcome, identifying an association, allocating attribution credit and estimating a causal or incremental effect are different kinds of conclusion. A separate question is how strongly the research design supports any causal interpretation being made.

For causal questions, randomized assignment generally provides the strongest design-based evidence because it creates variation in advertising treatment that is independent, in expectation, of consumers' prior propensity to respond. Well-designed quasi-experiments can also support causal inference where randomization is unavailable, but their credibility depends on additional assumptions about the construction of the counterfactual. Observational attribution, exposed-versus-unexposed comparisons and predictive models can provide valuable evidence about pathways, associations, heterogeneity and likely response, but do not establish causal effects by themselves.

This should not be interpreted as a permanent ordinal ranking of every named methodology. The quality of the individual study still matters. Statistical power, treatment integrity, outcome measurement, contamination, model specification, the credibility of the counterfactual and the scope of the estimand can all affect the strength of the resulting evidence. Scale affects precision and sometimes generalizability, but it is not itself the source of causal identification. A large observational dataset does not become experimental evidence because it is large, and a randomized design does not become informative if the treatment contrast or statistical power is inadequate.

For causal claims, evidence should be ranked by the strength of causal identification rather than by a universal hierarchy of methodological usefulness. Attribution may be more useful than an experiment for understanding observable pathways. MMM may be more useful for examining the total business and longer periods. Brand research may illuminate mechanisms that transaction data cannot observe. But when the question is whether changing advertising caused an outcome to change, evidence should be weighted according to how credibly the research design identifies that causal effect.

Exhibit: A hierarchy of evidence for causal advertising claims¹⁸

Evidentiary position	Illustrative approaches	What the evidence can support	Principal qualification
Strongest design-based causal evidence	Randomized user-, household-, device-, platform-, store- or geographic experiments; replicated evidence across comparable randomized tests	Causal effect of the defined intervention that was randomized on the population, outcome and period studied	Power, treatment integrity, contamination and outcome measurement still matter; the estimand may be bounded and may not generalize automatically
Conditional causal evidence under additional identification assumptions	Well-designed natural and quasi-experiments; difference-in-differences; synthetic controls; matched-market designs; Bayesian structural time-series approaches; appropriately specified causal models or MMM	Causal estimates where the assumptions required to construct the counterfactual are credible and adequately tested	Results depend more heavily on assumptions about comparability, trends, donor pools, specification, priors and unobserved influences
Observational, predictive and attribution evidence	Path-based attribution and MTA; exposed-versus-unexposed comparisons; observational matching and adjustment; predictive machine learning	Description of pathways, associations, prediction, operational credit allocation, heterogeneity and hypothesis generation	Selection and unobserved confounding can remain; these approaches do not establish incrementality by themselves
Integrated experimental-model evidence	Experiment-informed attribution; experimentally calibrated or constrained MMM; continuously validated causal models	Extends learning beyond the interventions directly randomized while allowing scalable models to be tested against stronger causal evidence	Model outputs inherit causal credibility only to the extent that the experimental evidence validly generalizes; uncertainty and domain boundaries must remain explicit

Note: The hierarchy describes general evidentiary strength for causal claims, not overall methodological usefulness, speed, scalability or granularity. Study quality and the scope of the question remain material within every category.

The two hierarchies answer different questions. The hierarchy of claims asks what kind of conclusion is being asserted; the hierarchy of evidence asks how convincingly a causal or incremental conclusion has been identified. A sophisticated model does not become causal simply because it produces a causal-looking number; equally, a randomized experiment does not support claims beyond the intervention, population, outcome and period actually tested. Good measurement requires alignment between the claim being made and the evidence used to support it.

Accumulating evidence across repeated experiments can strengthen confidence that findings generalize beyond a single test. Experiments can also be used to calibrate, constrain and validate scalable attribution or MMM systems. Such combinations are best understood as an evidence architecture rather than an additional rung in the hierarchy: models can extend causal learning beyond the campaigns directly randomized, but their causal credibility depends on how well the experimental evidence generalizes to the circumstances in which the model is applied.

¹⁸ Conceptual framing informed by Rick Bruner, Central Control, *Hierarchy of Advertising Evidence* (2025).

Why observational attribution can mislead

Observational attribution uses data generated through the normal operation of advertising campaigns, platforms, media systems and consumer behavior. It may connect exposures with subsequent purchases, compare exposed and unexposed consumers, analyze sequences of touchpoints or use statistical and machine-learning models to allocate credit among them.

Observational attribution remains valuable for describing consumer pathways, identifying differences in response, connecting campaign delivery with commercial outcomes and producing rapid signals for optimization. Its central limitation is the fact that advertising exposure does not normally occur at random.

Advertisers deliberately select audiences. Platforms optimize delivery toward consumers expected to respond. Consumers themselves enter media, search and commercial environments according to their interests and intentions. Television viewing, geography, demographics and retail access also affect both the likelihood of encountering advertising and the likelihood of purchasing.

As a result, exposed and unexposed consumers may have been different before the campaign began.

Conceptually:

$$\text{Observed difference in outcomes} = \text{advertising effect} + \text{pre-existing differences between exposed and unexposed groups}$$

The difficult task is separating these two components. Statistical methods can improve the comparison. Regression, matching, weighting, propensity scores and machine learning can make treatment and comparison populations more similar on recorded characteristics. But they can adjust only for variables that have been observed and represented adequately in the data.

They cannot guarantee that exposed and unexposed consumers are equivalent in unobserved purchase intent, brand preference, underlying need, previous advertising exposure or other latent characteristics.

A sophisticated model can become extremely good at predicting who will buy without becoming equally good at determining who bought because of the advertising.

Large-scale comparisons between observational methods and randomized experiments illustrate the problem. Research using 15 Facebook advertising experiments found that observational estimates based on extensive demographic and behavioral information frequently differed substantially from the causal effects measured through random assignment; for purchase outcomes, observational and experimental estimates differed by a factor of three or more in half of the studies¹⁹. Subsequent research using 663 experiments and more than 5,000 consumer-level variables found that some observational methods performed better than others, but none reliably reproduced experimental estimates across campaigns²⁰.

The implication is not that observational measurement always overstates effectiveness. Bias can operate in either direction. The important conclusion is that more data and more sophisticated prediction do not by themselves eliminate selection bias.

¹⁹ Source: Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava and Dan Chapsky, A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook (2019).

²⁰ Source: Brett R. Gordon, Robert Moakler and Florian Zettelmeyer, *Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement* (2023).

Targeting separates purchase propensity from advertising responsiveness

Targeting is designed to identify consumers who are likely to be valuable. This can increase observed conversion rates and reduce delivery waste.

But targeting creates a fundamental distinction between two questions:

Who is most likely to purchase?

and:

Whose probability of purchasing is most likely to change because they see the advertising?

These populations are not necessarily the same. A loyal customer may have an extremely high probability of buying but relatively little scope for the current campaign to change that behavior. A lapsed customer may be less likely to purchase overall but considerably more responsive to advertising.

This distinction can be expressed as:

Purchase propensity = probability of purchase without advertising.

Advertising responsiveness = change in purchase probability caused by advertising.

An optimization system trained to predict conversions may naturally concentrate advertising among high-propensity consumers. If performance is subsequently assessed by counting purchases among those consumers, the campaign can appear highly effective even when many of those purchases would have occurred anyway.

In other words: The better advertising systems become at finding likely buyers, the more important it becomes to distinguish prediction from incrementality.

Previous purchase behavior belongs at the center of the counterfactual

Previous purchase history is particularly important because it is sometimes treated simply as another audience characteristic or modeling variable.

In many categories, it is much more important than that. Past purchasing provides some of the strongest available evidence about what a consumer or household might have done without the current advertising. It can reveal category participation, purchase frequency, customer tenure, brand loyalty, repertoire behavior, likely purchase timing and propensity to switch.

Consider four households that purchase after exposure to the same campaign:

1. A loyal household has bought the brand every month for several years.
2. An occasional buyer alternates between the brand and competitors.
3. A lapsed buyer has not purchased the brand for a year.
4. A household has no recorded previous purchase of the brand.

The observable sequence is identical:

Advertising exposure → purchase

But the counterfactual is clearly different.

The loyal household may have been highly likely to buy anyway. The lapsed or new-to-brand household may have had a much lower baseline probability of purchasing. A more credible measurement question is: **Given this consumer's previous category and brand behavior, how likely was the outcome to occur without the advertising?**

This becomes even more important when previous purchases are used to select the advertising audience:

Previous purchase behavior → probability of advertising exposure

and:

Previous purchase behavior → probability of future purchase

The same variable then helps determine both treatment and outcome. Unless the analysis accounts for this relationship, baseline demand can easily be interpreted as advertising response.

Purchase history also changes the meaning of the outcome itself. A post-exposure purchase could represent:

- Acquisition of a genuinely new customer.
- Reactivation of a lapsed customer.
- Increased purchase frequency.
- Switching from a competitor.
- Retention of a customer who might otherwise have defected.
- Acceleration of a purchase that would have occurred later.
- Simple continuation of an established purchasing pattern.

A single attributed-sales figure collapses these economically different effects into one number.

Previous purchase behavior does not reveal the counterfactual perfectly. Recorded history may be incomplete, consumer circumstances change and purchase intention can develop rapidly. But where the data exist, it should generally be treated as **fundamental evidence about baseline demand**, rather than simply another optional model feature.

Other forms of observational distortion remain

Purchase propensity is not the only source of observational distortion. The principal additional risks – and the safeguards they imply – are summarized below. Connecting more datasets can improve observability without eliminating these identification problems.

Exhibit: Where observational attribution can go wrong

Source of distortion	Potential misinterpretation	Important safeguard
Selection bias	Differences between exposed and unexposed consumers are attributed to advertising.	Explain how exposure was generated and seek independent variation in exposure.
Pre-existing purchase propensity	Baseline demand is counted as campaign-created sales.	Use pre-campaign brand and category behavior to inform the counterfactual.
Targeting and optimization	High conversion rates are interpreted as high incrementality.	Distinguish purchase propensity from advertising responsiveness.
Missing variables	Intent, preference, price or other influences are assigned to advertising.	Use richer longitudinal data, sensitivity testing and experimental validation.
Cross-channel influence	Channels closest to conversion receive excessive credit.	Examine channel interactions and avoid adding platform-attributed outcomes.
Identity and exposure error	Household or device matches are treated as person-level causation.	Disclose the unit of exposure, identity method, coverage and uncertainty.
Incomplete exposure data	Observable touchpoints absorb the effect of unobserved activity.	State the measurement boundary and test alternative exposure definitions.
Attribution-window choice	Timing rules determine credit more than consumer response.	Test alternative windows and allow for carryover where relevant.
Predictive validation	A model that predicts buyers accurately is assumed to estimate causal effects accurately.	Validate causal estimates against experimental or credible quasi-experimental evidence.

What observational evidence must establish before stronger claims are made

Observational attribution remains useful, but the strength of the claim should depend on the strength of the design. Several disciplines can materially improve its credibility.

1. **Describe how advertising exposure was generated.** Measurement should explain audience eligibility, targeting rules, use of previous purchase behavior, platform optimization and why some eligible consumers received more advertising than others.
2. **Establish pre-campaign comparability.** Where possible, comparison groups should be assessed using previous category and brand purchasing, recency and frequency, customer tenure, previous media exposure and relevant behavioral signals.
3. **Report meaningful customer states separately.** New, lapsed, occasional and loyal customers can have radically different baseline purchase probabilities and advertising responses. Campaign averages can conceal these differences.
4. **Establish whether genuinely comparable unexposed consumers exist.** Statistical models become fragile when there is little overlap between exposed and comparison groups. If nearly every high-propensity consumer receives the advertising, the model must extrapolate what similar unexposed consumers would have done because few such consumers exist in the observed data.
5. **Test sensitivity.** Credible analyses should examine alternative attribution windows, exposure definitions, matching rules, historical lookback periods and model specifications. Conclusions that change materially under plausible assumptions should be reported as uncertain.
6. **Separate predictive from causal validation.** The ability to predict observed conversions in a holdout dataset is useful evidence about prediction. It does not establish that attribution weights correspond to incremental advertising effects.
7. **Validate against experiments wherever feasible.** Randomized or strong quasi-experimental evidence can provide independent benchmarks against which observational models can be tested.
8. **Match the precision of the claim to the identification strategy.** A household-level or campaign-level causal result should not be transformed automatically into person-level or impression-level causal certainty merely because the input data permit that resolution.

The principle is straightforward: **Granularity should be earned by the evidence, not inferred from the resolution of the data.**

Experiments provide a stronger counterfactual

Where randomization determines eligibility for advertising rather than guaranteeing actual exposure, the cleanest experimental estimate is generally the effect of assignment or eligibility – the intent-to-treat effect. Estimating the causal effect of actual exposure requires additional assumptions or analytical steps because consumers who comply with treatment can differ from those who do not. Experimental reporting should distinguish clearly between what was randomized and what was subsequently observed.

Experiments address the central weakness of observational attribution by deliberately creating variation in advertising eligibility that is independent of consumers' prior propensity to respond.

In a randomized experiment, eligible consumers, households or markets are assigned to different advertising conditions before outcomes are observed. With sufficient sample sizes and correct implementation, randomization makes treatment and control groups probabilistically comparable not only on recorded characteristics but also on unobserved factors such as prior intent or preference.

The resulting question is: **Did the defined advertising intervention cause an incremental change in this outcome, for this population, during this period?**

This is why randomized experiments provide the strongest available evidence for many advertising-effectiveness questions. Their value lies in answering the question actually randomized. If a campaign is randomized as a package, the experiment can estimate the incremental effect of that package. It should not be judged against a different question – such as the causal contribution of every impression, publisher, sequence or creative – unless the design created independent causal variation at that level.

Creative testing illustrates the same principle. A campaign holdout can identify the incremental effect of the campaign as implemented, including its creative strategy. Random assignment among creative executions can identify the incremental difference between defined executions. Observational differences among creatives may remain informative, but they should not inherit causal status merely because another aspect of the campaign was randomized.

This becomes especially important in complex consumer journeys. Attempting to experimentally isolate every combination of platform, sequence, frequency and exposure would rapidly create an infeasible number of treatment cells. In these circumstances, the appropriate objective may be to establish the incremental effect of a campaign, an additional channel, a different audience strategy or a change in spending, and use other evidence to investigate how that effect arose.

Different designs create different counterfactuals

The term *experiment* covers several different designs, each appropriate to different media environments.

Exhibit: Experimental and quasi-experimental approaches to incrementality

Approach	Basic design	Where it is particularly useful	Principal limitation
Individual-level RCT	Eligible individuals are randomly assigned to advertising eligibility or holdout.	Addressable environments with stable person-level identity and outcomes.	Cross-device exposure, identity loss and spillovers can contaminate assignment.
Household or device randomization	Households or devices are assigned to treatment or control.	CTV, addressable television and household-level commercial outcomes.	Does not necessarily identify which individual saw the advertising or acted.
Platform or campaign holdout	A platform withholds advertising from a randomly selected share of eligible users.	Measuring incremental effects inside a defined platform environment.	Establishes the effect of the platform-controlled intervention, not necessarily its contribution to the total media system.
Geo experiment	Markets or geographic areas receive different levels or combinations of media.	Television, radio, out-of-home, retail and integrated media activity that cannot be randomized easily at user level.	Few comparable markets, spillovers and local shocks can produce wide uncertainty.
Synthetic control / matched market	A treated market is compared with a weighted or matched set of untreated markets.	Situations in which deliberate randomization is unavailable.	Causal validity depends on assumptions about the stability and quality of the constructed counterfactual.

In person- or household-level randomized tests, an important distinction is between assignment and actual exposure. Consumers assigned to treatment are eligible to receive advertising; not all necessarily do so. Consumers in a proper control group are withheld from the focal treatment.

Comparing everyone actually exposed with everyone apparently unexposed can destroy the protection created by randomization because the delivery system may itself select who receives impressions. The causal comparison should normally preserve the randomized assignment.

The same principle applies to the level of inference. A household-randomized CTV test can provide strong household-level evidence without establishing which person within the household watched or purchased. A geo experiment can establish a market-level effect without allocating that effect deterministically to individual impressions. The unit of the causal claim should follow the unit of the experiment.

What can weaken an experiment and what falls outside its scope

Random assignment can create a credible causal comparison, but experimental evaluation requires separating threats to that comparison from questions the experiment was never designed to answer. Both affect how a result should be used, but they have different implications for the strength of the evidence.

What can weaken an experiment

Several problems can reduce confidence that an experiment has identified the causal effect it was intended to estimate:

- **Insufficient statistical power.** Advertising effects are often small relative to normal variation in commercial outcomes. Large samples or sufficient numbers of independent experimental units may be required, particularly for infrequent purchases, narrow audiences or relatively small changes in media activity. An inconclusive result is not necessarily evidence of no effect; the experiment may simply have been unable to distinguish the effect from noise.
- **Weak or inappropriate randomization.** The protection offered by an RCT depends on treatment assignment being genuinely randomized at an appropriate unit. Problems in assignment, post-randomization exclusions or analytical decisions that break the original treatment comparison can reintroduce systematic differences between treatment and control groups.
- **Insufficient treatment contrast.** The advertising conditions must differ enough for the experiment to test a meaningful intervention. If treatment and control groups receive very similar levels of the focal advertising, the experiment may provide little information about the effect of changing media activity.
- **Non-compliance and treatment leakage.** Consumers assigned to treatment may not actually receive the advertising, while consumers assigned to control may encounter it through another account, device, platform or media environment. These issues do not necessarily invalidate an intent-to-treat estimate, but they can weaken the treatment contrast and affect how the result should be interpreted.
- **Spillovers and contamination.** Advertising effects may cross experimental boundaries. Consumers can travel between geographic markets, share devices or households, influence one another through word of mouth or encounter related campaign activity outside the experimental environment. Where these effects are material, outcomes in one experimental condition may partly reflect treatment assigned elsewhere.
- **Weaknesses in outcome measurement.** Randomization does not correct errors in the outcome itself. Incomplete transaction coverage, inconsistent measurement across treatment and control groups, changes in matching rates or material differences in the population represented by the outcome data can weaken the resulting estimate.
- **Operational constraints on the design.** Creating a meaningful control condition may require advertisers to withhold advertising from part of an eligible audience or vary investment across markets. Commercial or operational constraints can lead to holdouts that are too small, tests that are too short or interventions that differ too little to answer the intended question reliably.

These are genuine issues of experimental design, implementation or measurement. They affect whether the comparison supports the causal conclusion being claimed and should be assessed explicitly.

What a valid experiment may still not establish

A well-designed experiment can provide strong causal evidence while remaining deliberately bounded. Its result applies to the intervention that was changed, the population and outcome studied, and the conditions under which the experiment was conducted.

The MRC's *Outcomes and Data Quality Standards* similarly require test-and-control results to be treated as estimates and emphasize treatment definition, exposure validation, contamination, comparison-group equivalence, statistical adequacy and bias assessment, providing an important existing foundation for this work.

A valid experiment may leave other questions unanswered:

- **The contribution of individual components.** If an integrated campaign or media package is randomized as a whole, the experiment can identify the incremental effect of that package without establishing the separate causal contribution of every publisher, platform, audience, creative execution or impression within it.
- **A different spending level or intervention.** A test comparing one level of investment with another estimates the effect of that particular change. It does not establish that the same return would apply at materially higher or lower spending levels or under a different media strategy.
- **A different population.** An effect identified among the consumers, households or markets included in the experiment should not automatically be assumed to apply to audiences with different characteristics or baseline propensities.
- **A different creative or campaign configuration.** A campaign-level experiment estimates the effect of the advertising treatment as implemented. If creative, targeting, frequency or other campaign characteristics change, the resulting effect may also change.
- **Effects beyond the measured outcome boundary.** An experiment using a particular transaction, retailer, platform or customer dataset can provide a credible estimate for outcomes captured within that measurement boundary while missing effects occurring elsewhere.
- **Effects beyond the observation period.** Advertising may produce delayed, cumulative or post-treatment effects. A short experiment can estimate near-term outcomes credibly without establishing the full longer-term effect of the advertising.
- **Generalization to different market conditions.** Results can depend on category conditions, seasonality, competitor activity, media prices, product availability and other features of the market in which the test occurred. A valid result from one setting should not automatically become a permanent estimate of a channel's effectiveness.
- **The mechanism through which the effect occurred.** An experiment may establish that changing advertising changed an outcome without revealing precisely why the effect occurred or which intermediate consumer responses produced it.

These are **boundaries on the scope and generalizability of the causal estimate, not failures of causal identification.** An experiment does not need to reconstruct every consumer pathway or decompose every component of a campaign in order to provide strong evidence about the intervention it was designed to test.

The distinction is important for interpreting experimental evidence. **A problem with randomization, treatment integrity, statistical power or outcome measurement can weaken whether the experiment establishes its intended causal effect. A question about another audience, component, spending level, outcome or period may simply lie outside the experiment's scope.**

Experimental results should be judged first on whether the design credibly identifies the causal effect it set out to estimate, and then on how far that estimate can reasonably be generalized. The fact that a randomized experiment produces a less granular or more bounded answer than an observational attribution system does not place the two on an equivalent causal footing.

The most important opportunity is to use experiments to calibrate attribution

The most important role for experiments in scalable attribution may be to anchor models that operate at levels of detail experiments cannot economically estimate for every campaign. Experimental evidence can identify incremental effects for defined interventions; observational models can then test how far those effects generalize across campaigns, audiences and touchpoints.

This approach is already beginning to appear in advanced attribution systems. Amazon's published multi-touch attribution methodology provides an important illustration. Amazon explicitly distinguishes between machine-learning models trained on observational data – which can scale and generate precise predictions but may yield biased estimates of advertising effects – and randomized controlled trials, which provide much stronger causal estimates but can be noisy and insufficiently granular. Its methodology uses experimental results to calibrate attribution models and then extends those causal signals toward more granular touchpoint-level allocation²¹.

The conceptual shift is important. The traditional assumption was often:

More exposure data + more outcome data + more sophisticated modeling = better attribution

The emerging model is:

Exposure and outcome data + observational models + experiments = better-constrained attribution

Amazon describes maintaining a large body of RCT evidence that feeds causal calibration models, which are then used to extend experimental signals to campaigns that have not themselves been randomized. The resulting attribution outputs can subsequently feed optimization systems²².

This does not mean that one experimental correction factor can be applied universally. Experimental effects remain specific to the intervention, population, period, outcome and spending conditions tested. Calibration becomes credible only where the model explains how experimental evidence is being generalized and where the experiments are sufficiently representative of the activity to which the model is applied.

But the underlying architecture is highly significant: **observational attribution is no longer required to prove causation from its own data alone.**

Experiments can also calibrate broader measurement systems

The same logic applies beyond multi-touch attribution. Marketing-mix models can estimate the effects of multiple channels, external demand drivers and longer periods, but their results depend on model specification and the historical variation available. Experiments can test whether a model recovers the effect of a known intervention, constrain implausible parameter estimates and inform prior assumptions.

Likewise, attention, search, website visits and other intermediate signals can be useful inputs into optimization, but their relationship with incremental business outcomes is not fixed. Experiments can establish whether changes in these signals are actually associated with causal improvements in sales, brand measures or customer value.

²¹ Source: Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero and German Schnaidt, *Amazon Ads Multi-Touch Attribution* (August 2025).

²² Ibid.

The same principle becomes more important as measurement feeds automated decision systems: experimental evidence can test whether the signals being optimized correspond to the incremental business outcomes advertisers actually value. Over time, repeated tests can provide a body of causal evidence against which scalable models are recalibrated and revalidated.

From attribution to a portfolio of evidence

The central lesson is not that observational attribution should be discarded or that every marketing decision requires an RCT. Attribution, experiments, quasi-experiments, MMM and longitudinal customer evidence answer different questions. The strongest architecture is one in which stronger causal evidence constrains more scalable but assumption-dependent inference.

This has an important implication for the meaning of attribution itself. Attribution should be understood not as the mechanical allocation of a known outcome across preceding exposures, but as an attempt to estimate **relative incremental contribution**.

That changes the standard against which attribution models should be judged.

Attribution models should be judged less by the number of impressions to which they can allocate credit than by whether their estimates remain consistent with credible causal evidence when such evidence becomes available.

The distinction leads back to three questions:

1. **Who bought after seeing the advertising?**
2. **Who was already likely to buy?**
3. **Whose behavior changed because of the advertising?**

The first is an observational question.

The second is a counterfactual modeling question.

The third concerns causal contribution. A stronger attribution system moves from observed response toward causal contribution while making the uncertainty between those two quantities explicit.

Stronger causal identification addresses a central part of the measurement-confidence gap: it improves confidence that a defined change in advertising caused a defined change in an outcome. But it does not solve every measurement problem.

Even a well-designed experiment is necessarily bounded by the outcome, intervention, population and period it measures. If the outcome is a near-term purchase, the experiment can establish that purchase incrementality credibly while still missing advertising effects expressed through memory, consideration, retention, switching, future category entry or longer-term customer value. Strong causal evidence about a narrow outcome is not the same as a complete measure of advertising effectiveness.

Stronger causal identification solves only part of the measurement problem. Even a well-designed experiment is bounded by the outcome, intervention, population and period it measures. Where advertising creates value through memory, consideration, retention, switching, future category entry or longer-term customer value, a credible near-term causal estimate may still represent only part of the effect that matters.

Chapter 8:

Beyond the next transaction – measuring brand and long-term effects

The most readily measured outcomes may capture only part of advertising's value. Stronger experiments and credible quasi-experiments can establish more convincingly whether a defined advertising intervention changed a defined outcome. But stronger causal design does not by itself solve a second problem: the outcome being measured may capture only part of the value advertising creates.

Brand-building and long-term effects often emerge beyond the next observable transaction. Advertising can prompt an immediate response, but it can also influence memory, consideration, retention, switching, future category entry and demand over weeks, months or years. Systems optimized around near-term transactions can measure an important part of advertising's impact while under-representing effects that create, reinforce or preserve future demand.

Individual transaction attribution – even when causally strengthened – cannot by itself measure the full value of advertising because many effects are delayed, cumulative, probabilistic, preventative or expressed across future purchase occasions.

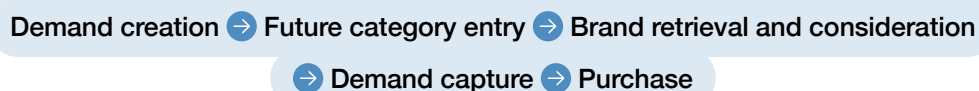
Demand capture and demand creation are different functions

Advertising can help capture demand that already exists. It can also create, sustain and shape demand before consumers are actively in the market.

Demand capture occurs when a consumer already has a developed need or intention and advertising helps direct that demand toward a particular brand or seller. Search advertising, retail media, retargeting and promotional activity can be particularly effective in this role. Because these interventions sit close to purchase, their relationship with subsequent behavior is also comparatively easy to observe.

Demand creation occurs earlier and is less visible. Advertising may establish category associations, build familiarity, encode distinctive brand assets, communicate reasons to believe or simply make a brand more likely to come to mind when a future need arises. The consumer may not currently be shopping and may generate no observable behavioral response during the campaign.

A simplified pathway looks like this:



Demand creation and demand capture are not competing theories of advertising. They are complementary functions within the same system. Activity close to the transaction can help convert existing demand efficiently; earlier activity can increase the volume, composition or resilience of the demand subsequently available to be captured.

The attribution problem is that the final measurable touchpoint usually sits much closer to the transaction. A consumer may see television advertising, encounter the brand repeatedly over several months, later enter the category, search for the brand and purchase through a retailer. A path-based attribution system may assign substantial credit to search or retail media because they are visible immediately before the transaction.

That may accurately describe the final observable pathway. It does not establish where the underlying demand came from. Proximity to the transaction is not the same as importance in producing the transaction.

This distinction becomes particularly important when measurement systems are used to optimize media. A system that consistently rewards the channels best positioned to capture existing demand can gradually redirect investment away from activities whose contribution is to create the future demand those channels subsequently harvest.

Creative determines what the exposure can do

Media creates the opportunity for advertising to work; creative determines what the consumer encounters within that opportunity.

An impression is not a standardized unit of advertising effect. Two impressions delivered to otherwise comparable consumers in comparable environments can produce different responses because the advertising itself differs. Creative can affect whether an exposure gains attention, evokes emotion, creates or refreshes memory, communicates the brand, changes perceptions or prompts action. These mechanisms sit between media delivery and many of the intermediate and commercial outcomes examined in this paper.

The distinction applies to both demand creation and demand capture. Creative intended to create future demand may seek to establish familiarity, strengthen distinctive brand assets, build associations or make a brand more likely to be retrieved when a future need arises. Creative closer to the transaction may communicate a product, offer, reason to act or other information intended to influence an immediate decision. The appropriate measure of creative effectiveness consequently depends partly on what the advertising is intended to change and when that change is expected to become visible.

A substantial body of advertising research indicates that differences in creative execution are an important source of variation in campaign effectiveness and can rival or exceed differences associated with audience targeting or media delivery. The precise contribution varies by campaign, category, objective, methodology and outcome, but treating creative as a constant can materially distort assessments of advertising performance. Attention, emotion, memory, motivation and persuasion are distinct but interconnected processes through which advertising can influence consumer response. Creative execution can affect several of these processes, while their relevance varies according to the advertisement, audience, context and objective. Research examining individual creative attributes similarly finds that different executional characteristics can be associated with different effects on recall, awareness, consideration, favorability and purchase intent. Creative effectiveness is multidimensional rather than reducible to whether an advertisement simply “worked” or “did not work”²³.

Creative variation also creates a measurement problem. If different publishers, audiences or periods receive materially different executions, some apparent media or audience effects may reflect differences in the advertising stimulus itself. Dynamic creative can intensify the problem because the execution received may be selected using audience characteristics, prior behavior or predicted response. Observed creative performance can combine creative effects with selection effects.

23 Sources: ARF Cognition Council Cognitive Processes Working Group, *The Brain on Advertising: Why We Notice, Feel, Remember and Act on Ads* (May 2026); Ryan S. Elder, Gina Slejko, Jeff Dotson and Ariane Pol, *Designing Delicious: An Examination of Creative Attributes Driving Food, Beverage, and Restaurant Advertising Effectiveness* (2025).

Where the decision concerns creative, the research design should create the relevant causal variation. Random assignment among executions can estimate differences between defined creative treatments; a campaign-level holdout can estimate the effect of the campaign as implemented without decomposing that effect across individual executions. Attribution and MMM should represent material creative variation where possible, but creative-level granularity should not be interpreted as creative-level causality without an appropriate counterfactual.

The objective is not a universal creative-quality score. It is to make meaningful variation in the advertising treatment visible and to use measures suited to the response the creative is intended to produce – whether immediate action, attention, memory, brand association, consideration or longer-term demand.

Mental availability creates value before a transaction occurs

One important mechanism through which advertising can affect future demand is mental availability: the probability that a brand will be noticed, recognized or retrieved in a relevant buying situation.

Consumers do not generally evaluate every possible brand before making a decision. They operate with limited attention, incomplete information and imperfect memory. Brands that come readily to mind – or are recognized easily when encountered – have an advantage when a purchase occasion arises.

Advertising can contribute by building and refreshing memory structures and linking the brand with the needs, situations or occasions that bring consumers into the category.

This creates several measurement challenges:

The effect is **probabilistic**. Advertising does not need to make every exposed consumer purchase to create value. A small increase in the probability that a brand is retrieved or chosen, distributed across a large population and many future purchase occasions, can be commercially important while remaining difficult to identify at individual level.

Memory is both cumulative and perishable. Advertising can build associations over repeated exposures and reinforce them over time. It can also prevent existing associations from weakening. The commercial outcome may be the maintenance of an existing advantage rather than an immediate uplift from one campaign.

The effect may not become commercially observable until the consumer next enters the category. That could occur within days for frequently purchased products, but months or years later for automobiles, financial services or other low-frequency categories.

A conventional attribution window necessarily struggles with this mechanism. A 7-, 14- or 30-day window may be operationally convenient, but it is not a theory of how long advertising effects actually last.

Brand research can help by measuring nearer-term changes in awareness, consideration, memory, distinctive-asset recognition or category associations. But these metrics should reflect how the advertising is intended to work. Mental availability is not synonymous with awareness, and no single brand metric provides a complete measure of future commercial value.

An established brand trying to broaden the situations in which it comes to mind should not necessarily be evaluated using the same metric as a new brand trying to establish basic awareness.

The appropriate question is: What change in the consumer is the advertising intended to produce before the eventual transaction occurs?

Baseline demand is not static

A second important implication follows from the concept of baseline demand. Every established brand begins a campaign with sales that would have occurred without the current advertising. Those sales reflect many influences: product experience, price, distribution, reputation, word of mouth, previous advertising and accumulated customer relationships.

For short-term attribution, this baseline is principally something that measurement must distinguish from the incremental effect of the current campaign.

For long-term effectiveness, however, baseline demand is also partly an accumulated asset.

Conceptually:

$$\text{Current sales} = \text{Baseline demand} + \text{Current marketing effects} + \text{Other current influences}$$

But baseline demand itself changes over time:

$$\begin{aligned} \text{Future baseline demand} &= \text{Current baseline demand} + \text{Effects carried forward from current activity} \\ &\quad - \text{Decay and competitive erosion} \end{aligned}$$

Advertising today can contribute to the level of demand from which future periods begin. This produces an asymmetry in conventional measurement: advertising cost is recognized immediately, while part of the return may emerge later. The cost of advertising is recognized immediately and completely. Part of its return may emerge much later.

A reduction in brand investment can consequently appear efficient in the short term because current sales remain supported by memory, existing customers, distribution and demand accumulated through previous activity. The deterioration, if it occurs, may become apparent only later as familiarity weakens, customer flows change or competitors gain ground.

Similarly, current performance marketing can appear extremely productive because it is converting demand partly created through earlier investment.

The measurement implication is important: **Baseline demand should not always be treated as though it arose independently of historical marketing.**

This does not mean that every future sale should be credited retrospectively to past advertising. Product quality, price, availability, customer experience and many other factors also shape baseline demand. It means that long-term measurement requires a dynamic rather than static view of the counterfactual.

Recent effectiveness analysis similarly emphasizes that marketing effects can persist well beyond the initial response period, making a short-term-only evaluation an incomplete representation of economic return²⁴.

²⁴ Source: Google, *The Effectiveness Equation* (2025).

Long-term effects cannot be represented by one generic decay curve

Advertising carryover is often modeled as though an immediate effect simply becomes progressively smaller over time. This can be a useful approximation, but different mechanisms can create very different temporal patterns.

Advertising may generate:

- **Immediate effects**, such as a prompt search, enquiry or purchase.
- **Delayed effects**, when response occurs only when a later need or purchase opportunity arises.
- **Cumulative effects**, where repeated activity is required before behavior changes.
- **Memory effects**, which build and decay over time.
- **Threshold effects**, where activity has little observable impact until sufficient reach or continuity is achieved.
- **Reinforcement effects**, where advertising increases the effectiveness of subsequent search, retail, promotional or sales activity.
- **Preventative effects**, where value is created through reduced churn, switching or decline.

These processes need not share the same timetable. A promotional response might disappear rapidly. A brand association may persist considerably longer. A consumer exposed to an automobile campaign might purchase months later because the category itself has a long buying cycle. A subscription campaign may create value through retention that becomes observable only at the next renewal point.

Longer measurement periods capture more potentially relevant effects but also introduce more competing influences. But as causal distance increases, so does the number of other influences that may affect the outcome – later media, competitor activity, price, promotions, economic conditions, availability and consumer circumstances.

There is no universally correct attribution window. **An attribution window is an assumption about the plausible duration of an effect, not a neutral characteristic of the data.**

Modern MMM and other time-series approaches can estimate more flexible carryover and saturation effects, but model flexibility does not eliminate the problem. The greater the freedom given to the model, the greater the need for sensible priors, experimental calibration, sensitivity testing and external validation.

Long-term value extends beyond the first conversion

For many businesses, the first transaction represents only a fraction of the economic value of the customer relationship. Customers may purchase repeatedly, remain subscribed, renew contracts, upgrade, refer other customers or become less costly to serve. Advertising may also affect the quality of customers acquired: their subsequent margin, purchase frequency, retention or responsiveness to future marketing.

Two campaigns producing identical numbers of initial conversions can create very different commercial value. Customer lifetime value provides a framework for considering these effects. But it introduces another distinction between **causation and valuation**.

A lifetime-value model may estimate that a newly acquired customer is expected to generate \$1,000 of future contribution. That does not mean the advertising created \$1,000 of incremental value. Some of those customers may have joined anyway.

A stronger framework separates two questions:

1. **How many customer relationships did the advertising incrementally create, preserve or change?**
2. **What is the expected economic value of those incremental relationships?**

The first is a causal measurement problem.

The second is a valuation problem.

Both contain uncertainty, but they are different uncertainties.

This distinction is particularly important in subscription services, financial services, telecommunications, travel and other categories in which retention and future behavior dominate the economics of customer acquisition. The same principle can apply to frequently purchased goods where advertising affects purchase frequency, repertoire behavior or retention rather than simply triggering one additional transaction.

Recent effectiveness work has similarly argued for looking beyond the initial conversion toward customer lifetime value and the effect of brand building on repeat behavior and future economic value²⁵.

Why transaction attribution is structurally incomplete

The limits of individual attribution extend beyond missing cookies, incomplete identity graphs or imperfect datasets.

There is also a more fundamental issue: **an individual advertising exposure linked with an individual transaction is not the natural unit in which every advertising effect occurs.**

Transaction attribution is likely to underrepresent value when:

- Advertising affects purchase occasions outside the observation period.
- Advertising changes purchase probability slightly across a large population rather than creating a conspicuous individual response.
- The principal effect is on memory or consideration before consumers enter the category.
- Advertising changes retention, switching or churn rather than generating an immediate acquisition.
- Advertising changes purchase frequency, order value or future margin.
- Media activity affects price realization, distribution or other market-level outcomes.
- Demand created by one channel is subsequently credited to a lower-funnel channel.
- The eventual outcome occurs through an identity or retail environment that cannot be connected to the original exposure.

At the same time, transaction attribution can overstate value when purchases that would have occurred anyway are assigned to recent advertising exposures.

The problem operates in both directions. A transaction system can overcredit some immediate purchases while simultaneously failing to capture some of advertising's longer-term effects.

²⁵ Ibid.

This is why improving identity resolution alone cannot create a complete account of advertising effectiveness. Even perfect observation of every exposure and every transaction would still leave the analyst with questions about causation, time, memory, future demand and effects that are not naturally expressed as discrete conversions.

Transaction attribution should consequently be viewed as a **bounded instrument**, not a complete balance sheet of advertising value.

Brand metrics provide evidence, but not financial proof

The obvious response to the limits of transaction data is to measure the intermediate outcomes through which advertising is expected to affect future behavior. Brand studies can test awareness, associations, recognition, consideration, preference, trust, intent and other responses. These measures can provide important evidence about whether advertising changed consumers in the intended direction.

Intermediate brand effects can provide important evidence about mechanism and future demand, but they should not acquire causal or financial status merely because they are measurable. A statistically significant change in awareness, consideration, preference or branded search can establish that an intermediate outcome changed; it does not, by itself, establish the commercial value of that change. Their commercial interpretation is strongest where longitudinal, experimental or aggregate evidence connects changes in those measures with subsequent behavioral or economic outcomes. Intermediate effects can be valuable evidence without being treated automatically as monetized outcomes.

Brand measurement is strongest when four disciplines are applied:

1. **The metric reflects the intended mechanism:** A campaign designed to build relevant memory associations should not be judged exclusively through immediate purchase intent.
2. **The comparison supports causal inference:** Where possible, exposed and control groups should be constructed experimentally or with a transparent account of pre-existing differences.
3. **The relationship with subsequent behavior is tested:** Where commercial interpretation is important, the analysis should examine whether changes in the intermediate measure are followed by, predict or form part of a credible causal pathway to subsequent behavioral or economic outcomes.
4. **Uncertainty is disclosed:** Sample size, confidence intervals, multiple metrics, exclusions and analytical choices matter just as much for brand studies as for transactional attribution.

The objective should not be to convert every movement in a brand metric immediately into dollars, but to establish a credible chain of evidence:

Advertising exposure → Intended intermediate effect → Subsequent behavior → Commercial value

Not every study will observe every stage of this chain. Measurement should make clear which stages have been observed and which remain inferred.

Long-term measurement requires several levels of evidence

No single methodology can provide a complete measure of long-term advertising effectiveness. The practical solution is to combine evidence operating at different levels.

Exhibit: A layered framework for measuring brand and long-term effects

Layer	Principal question	Illustrative evidence	Principal limitation
Exposure and attention	Did the advertising reach consumers and have an opportunity to be processed?	Reach, frequency, attention, completion, exposure quality.	Does not establish memory, behavior or causation.
Memory and brand response	Did advertising change the mental or attitudinal outcomes it was intended to influence?	Awareness, associations, category entry points, distinctive-asset recognition, consideration.	Intermediate effects do not automatically translate into commercial outcomes.
Near-term behavioral effects	Did a defined intervention change prompt actions or transactions?	RCTs, geo tests, quasi-experiments, search, visits, purchases.	May miss delayed, cumulative or preventative effects.
Customer development	Did advertising change acquisition quality, repeat purchasing or retention?	Cohort analysis, frequency, churn, margin, longitudinal holdouts, incremental lifetime value.	Requires longer observation periods and assumptions about future value.
Aggregate business effects	Did advertising change total demand and economic performance over time?	MMM, market experiments, sales, penetration, price, distribution, branded demand.	Aggregate and model-dependent; causal separation can remain difficult.

These layers should not be collapsed mechanically into a single number. A brand-lift experiment measuring immediate consideration, a six-week sales holdout and an MMM estimating a channel's contribution over several years are not necessarily contradictory if they produce different results. They are measuring different outcomes, over different periods and frequently at different levels of aggregation.

The purpose of triangulation is not to average them together but to determine whether the different pieces of evidence form a coherent account of how advertising affected the business.

A practical standard for long-term outcomes claims

Measurement design should begin with the claim. Any claim about brand or long-term advertising value should specify eight elements:

1. **What effect is being claimed?** Is the advertising expected to strengthen memory, increase penetration, accelerate purchase, reduce churn, support price or increase customer value?
2. **For whom?** Non-buyers, light buyers, existing customers, lapsed customers, category entrants or the market overall?
3. **Over what period?** Days, months, a purchase cycle, a renewal period or several years?
4. **Against what counterfactual?** What would probably have happened without the advertising, including natural decay, competitive activity and historic brand momentum?
5. **At what level has causality been identified?** Campaign, channel, market, household, audience or individual exposure?
6. **How is intermediate evidence connected with commercial value?** Through longitudinal evidence, experiments, historical relationships or modeling?
7. **What remains outside the measurement boundary?** Future purchase occasions, unmatched transactions, retention, pricing or wider market effects?
8. **How sensitive is the result to assumptions?** Attribution windows, lag structures, customer-value assumptions, model specification or priors?

These questions do not eliminate uncertainty. They make clear **where the uncertainty resides**.

The objective is a fuller account of how advertising creates value

Brand and long-term effects are not peripheral issues that can be considered after transaction attribution has been perfected. They are part of the central measurement problem.

A credible measurement system must avoid two opposite errors:

1. The first is **over-attribution**: treating purchases that occur after advertising exposure as though they were necessarily caused by the advertising. This risks assigning current campaigns credit for baseline demand, existing customer behavior and purchases that would have occurred anyway.
2. The second is **under-measurement**: recognizing only those advertising effects that can be connected quickly to an individual transaction. This risks excluding value created through memory, future category entry, increased purchase frequency, retention, reduced switching and other effects that emerge beyond the immediate attribution window.

These are not merely reporting errors. Once outcomes measures influence bidding, optimization and budget allocation, the measurement horizon can begin to shape the media strategy itself. A system that observes immediate conversion more reliably than future demand creation may systematically reward activity positioned close to an existing purchase while undervaluing advertising that changes memory, consideration, future category entry or customer value over a longer period. Measurement can create an optimization bias: the effects easiest to observe become the effects the system selects for.

The policy implication is not that short-term outcomes should be discounted. It is that the decision system should not silently equate measurability with value.

These are two different forms of measurement-confidence failure. Over-attribution assigns greater causal certainty than the evidence warrants; under-measurement defines the relevant outcome too narrowly and omits value that the advertising may genuinely have created.

The measurement requirement is twofold: distinguish current advertising effects from demand that would have existed anyway, while also extending the measurement horizon far enough to capture value created through memory, customer development and future demand. No single method can perform both tasks across every category and time horizon.

The full value of advertising is not contained in the next observable transaction. It lies in the difference advertising makes to the future path of demand.

A single-system view of outcomes measurement encounters three different limits. More granular data do not automatically provide causal confidence. Stronger causal designs remain bounded by the intervention and outcome tested. And even a credible short-term causal estimate may capture only part of the value advertising creates.

The remaining problem is architectural: how should granular attribution, causal experiments, MMM, brand evidence, customer data and other forms of measurement be combined so that each contributes what it can establish, stronger evidence constrains weaker inference, and no single output is asked to represent the whole effect?

Chapter 9:

From measurement systems to measurement intelligence

Next-generation measurement is better understood as a connected portfolio of evidence than as a single model. It is more likely to be a system that learns across models, a connected portfolio of evidence in which different methods operate at the levels and time horizons they can credibly support, test and constrain one another, and make their assumptions and remaining uncertainty explicit.

Attribution can provide speed, scale and operational granularity. Experiments can provide stronger causal anchors for defined interventions. MMM can place advertising within wider cross-channel and business dynamics over longer periods. Brand and customer evidence can capture mechanisms and effects that transaction data may not reveal. Identity and exposure measurement define what is observable in the first place.

The task of next-generation measurement is to make these forms of evidence operate as a coherent learning system rather than as competing products producing disconnected answers.

Artificial intelligence can reduce the cost and operational complexity of this architecture, but it is an enabling layer rather than a new category of measurement evidence. AI changes the cost, speed and accessibility of measurement more readily than it changes its epistemology. It can help integrate evidence arriving at different frequencies, operate complex models, automate diagnostics, identify inconsistencies, generate hypotheses and make probabilistic evidence easier for decision-makers to interrogate. It cannot create an unobserved counterfactual, repair a badly designed experiment, recover exposures that were never measured or transform association into causation simply by processing more data.

AI is most useful in making a continuously learning evidence system operational: integrating evidence, testing contradictions and sensitivity, updating models and translating uncertainty into decision support.

Responsibility for the evidence system is distributed across the marketplace. Advertisers ultimately own the business questions, decisions and standards of evidence against which measurement should be judged, often working through their agencies and internal analytics teams. Measurement providers are responsible for producing transparent, validated and appropriately bounded estimates. Platforms, publishers and data providers must make sufficiently robust exposure, outcome and experimental evidence available for those estimates to be tested. Industry bodies can establish the common definitions, disclosure requirements and validation frameworks that allow evidence from different systems to be compared. This architecture should be understood not as a single product, but as a set of responsibilities distributed across the measurement ecosystem.

The architectural question is how to assign each form of evidence a clear role and allow the evidence to accumulate without collapsing different claims, time horizons and levels of causal confidence into a single metric.

From measurement product to measurement system

Many measurement programs still operate as collections of largely independent products. A platform reports attributed conversions. An identity provider estimates reach. A research supplier measures brand lift. An analytics team runs an MMM. Occasional experiments are commissioned to answer particular questions.

Each output may be valuable. The problem is that they are frequently presented alongside one another rather than integrated into a coherent process of inference.

Different methods may use different definitions of exposure, different populations, different outcome windows and different counterfactuals. Their estimates can diverge substantially without it being clear whether the difference represents a genuine finding, a difference in scope or an error in one of the systems. Without an explicit understanding of the relative strength of different forms of evidence for causal claims, the most granular, immediate or commercially convenient number can acquire more authority than its identification strategy justifies.

A next-generation architecture would start somewhere different: with the decision to be made and the effect that needs to be estimated. Advertisers, working with their agencies, internal analytics teams and measurement partners as appropriate, should define the relevant outcome, population, time horizon and counterfactual before selecting the measurement method. The architecture can then bring together seven interdependent layers:

1. Identity and exposure measurement.
2. Creative treatment and response measurement.
3. Observational attribution.
4. Experimental evidence.
5. Modern marketing mix modeling.
6. Brand and intermediate-effect measurement.
7. Causal integration and decision modeling.

These are not seven competing routes to the same number. They observe different parts of the advertising and consumer-response system, at different levels of aggregation and over different periods. Their value comes partly from their individual findings, but increasingly from the ways those findings can inform, calibrate and challenge one another.

The hierarchy of evidence and the architecture of evidence solve different problems. The hierarchy concerns the relative weight that different evidence should carry when the claim itself is causal. The architecture concerns how evidence serving different purposes, operating at different levels of granularity and covering different time horizons can be combined into a coherent measurement system. The seven layers below should not be read as a ranking. Experimental evidence occupies a privileged position for causal identification, while attribution, MMM, brand, customer, creative and exposure evidence remain indispensable for scale, context, mechanism, diagnosis and decision-making.

Exhibit: A layered architecture for next-generation outcomes measurement

Layer	Principal question	Role in the architecture	Principal caution
1. Identity and exposure measurement	Who, or which households, had an opportunity to see the advertising, where and how often?	Defines campaign delivery, reach, frequency, duplication and the observable population.	Better matching does not establish causality; identity error and unrepresentative matching can bias subsequent analysis.
2. Creative treatment and response measurement	What advertising stimulus was delivered, and what evidence describes or predicts how different executions are likely to generate attention, emotion, memory, brand response or behavior?	Makes material variation in the advertising treatment visible and provides creative-level inputs to attribution, MMM, experimentation and brand analysis.	Creative-response measures may be directly observed or model-predicted; neither should automatically be interpreted as a causal creative effect. Performance can vary by audience, context, objective, time horizon and assignment mechanism.
3. Observational attribution	Which exposures and touchpoints are associated with subsequent response?	Provides fast, granular diagnostic signals and supports tactical optimization.	Prediction and association can be mistaken for incrementality.
4. Experimental evidence	What changed because of a defined advertising intervention?	Provides the strongest direct evidence of incrementality and calibrates other methods.	Experiments estimate defined treatments in defined populations; they are not universal ground truth.
5. Modern MMM	How do media and other business drivers jointly affect aggregate outcomes?	Provides the cross-channel, business-level and longer-term strategic spine.	More sophisticated modeling does not remove identification, specification or data limitations.
6. Brand and intermediate-effect measurement	How is advertising changing attention, memory, consideration, search and other precursors to demand?	Helps explain mechanisms and captures effects that may precede observable sales.	Intermediate effects should not automatically be treated as economic value.
7. Causal integration and decision modeling	What conclusions remain credible when the evidence is considered together?	Reconciles evidence, propagates uncertainty and supports decisions under risk.	AI can manufacture apparent coherence and precision where the underlying evidence remains weak.

What each layer contributes

The first layer establishes **what advertising was delivered and what can actually be observed**. First-party data, automatic content recognition, set-top-box data, publisher logs, retail transaction data and privacy-preserving environments can substantially expand the observable evidence base. Data clean rooms, for example, can support exposure-to-outcome analysis, test-and-control measurement, cross-publisher reach and frequency analysis and aggregated data feeds into MMM.

But this remains measurement infrastructure rather than proof of effectiveness. A clean room can create a controlled environment in which previously separated datasets can be analyzed together; it cannot convert correlation into causation. Identity quality, population coverage, device-to-person assumptions, differential match rates and missing exposures remain measurement variables whose uncertainty should be carried into subsequent analysis rather than concealed.

The second layer is creative treatment and response measurement. Exposure data establish whether and under what conditions an advertising opportunity was delivered; creative data describe the stimulus contained within that opportunity. Creative identifiers, asset-level linkage and structured executional metadata can make differences in message, branding, imagery, sound, format, offer and other creative characteristics visible to the wider measurement system.

Creative information can then become an input to attribution, experimentation, MMM and brand measurement rather than remaining an unobserved source of variation. Where particular executions are concentrated among particular audiences, publishers, placements or periods, representing creative variation can help distinguish differences in what consumers received from differences in where or to whom the advertising was delivered.

The evidentiary status of these inputs should remain explicit. Directly measured consumer response and model-predicted response are different forms of evidence, and a creative characteristic associated with stronger outcomes has not necessarily been shown to cause those outcomes. Creative measurement should make the advertising treatment more visible without implying either a universal creative-quality score or creative-level causality that the research design cannot support.

The third layer is observational attribution. It can provide scalable, rapid and granular evidence about observable pathways, creative and audience variation and changing campaign performance. Its causal status should remain bounded: exposure followed by purchase is not the same as an incremental purchase, and observational estimates should be tested against stronger causal evidence where the decision requires incrementality.

Its role should nevertheless be explicit. Exposure followed by purchase is not the same as an incremental purchase. Statistical adjustment is not necessarily causal identification. And assigning every conversion across recorded touchpoints does not demonstrate that those touchpoints collectively caused the conversion.

Prior purchase history can be central to the counterfactual. In many categories, previous behaviour is a major predictor of what would have happened without the advertising. A household with a long record of buying the brand may be highly likely to purchase after exposure while having little incremental response to the advertising itself. A sophisticated algorithm that cannot adequately observe prior purchase behaviour can produce a more intricate description of the wrong counterfactual.

The fourth layer is **experimental evidence**. Randomized experiments remain the strongest general design for determining whether a defined intervention caused an incremental change in outcomes and should consequently occupy a privileged position in the evidence hierarchy. A platform holdout, household experiment or geographic experiment estimates the effect of the intervention it randomized. The scope of that estimand should remain explicit when its results are used to calibrate broader models.

Experiments should function as causal anchors and calibration points. They are particularly valuable where observational bias is likely to be consequential: establishing whether targeted consumers were persuaded rather than simply selected; whether additional frequency is incremental; whether an addressable strategy outperforms broader delivery; or whether platform-reported conversions survive a randomized holdout. The fifth layer is **modern MMM**,

which provides the broadest business-level view. It can incorporate television, streaming, digital and retail media alongside price, promotions, distribution, seasonality, competitor activity and economic conditions. It does not depend on reconstructing every individual consumer journey and can capture longer-term carryover and saturation effects that short-window attribution may miss.

Modern Bayesian and more granular MMM approaches create additional opportunities. Experimental results can inform prior distributions or calibration ranges. Delivered reach and frequency can replace crude spending proxies. Brand measures can represent intermediate pathways. Models can update more frequently and express uncertainty directly.

But modernization does not repeal the fundamental problem of identification. Media spending can move with expected demand. Channels can be too closely correlated to estimate separately. Results can depend materially on priors, lag structures, transformations and control variables. More data may reduce sampling error without resolving the causal question.

The sixth layer incorporates **brand and intermediate effects**. A measurement system focused only on near-term conversions is structurally biased towards short, observable consumer journeys. Much advertising works by changing future probabilities: strengthening memory, increasing familiarity, building consideration, generating search, reducing perceived risk or making a brand easier to retrieve when a later purchasing occasion arises.

Measures of attention, brand linkage, awareness, consideration, preference, branded search, inquiries, store visits, switching, retention and price sensitivity can provide important evidence about how advertising works.

The central challenge is to connect these effects to behaviour rather than treating every movement in an intermediate metric as economically valuable. Experiments can establish whether advertising changes a brand measure; longitudinal analysis, MMM and causal models can then examine whether those changes predict later commercial outcomes. This is especially important for television and other broad-reach media whose contribution can be systematically understated when only responses close to purchase are counted.

The final layer is causal integration and decision modeling. Its purpose is not simply to average all the preceding estimates.

A stronger approach begins with an explicit model of the hypothesized causal relationships among advertising exposure, targeting, previous purchasing, brand effects, search, promotions, price, distribution, competitors and sales. Different evidence can then enter the system according to what it actually identifies.

Experimental evidence might inform plausible ranges for channel effects. MMM can connect those effects with business-wide outcomes and longer time horizons. Observational attribution can contribute information about timing and heterogeneity without being permitted to determine the causal total. Brand evidence can describe intermediate pathways. Identity-quality estimates can contribute measurement uncertainty.

Where evidence converges for genuinely independent reasons, confidence should increase. Where it diverges, the disagreement should itself become an object of analysis.

An attribution estimate may exceed an experimental estimate because it captures pre-existing purchase propensity. A short-term experiment may appear weaker than an MMM estimate because the latter incorporates carryover effects. A platform experiment may conflict with a national model because the measured platform population is systematically different from the total market.

A good architecture exposes these differences rather than forcing them into artificial agreement.

AI as an enabling layer

AI can strengthen the evidence architecture in four principal ways.

Integration: bringing together evidence generated by attribution, experiments, MMM, brand research, customer data and exposure measurement while preserving the differences in scope, timing and evidentiary strength among them.

Criticism: testing sensitivity, surfacing contradictions, identifying unstable assumptions and distinguishing genuine corroboration from several analyses that depend on the same underlying data or model.

Adaptation: updating models as new experiments, outcomes and market conditions become available, and identifying where additional evidence would reduce the most decision-relevant uncertainty.

Decision support: making complex statistical evidence more accessible, allowing decision-makers to interrogate assumptions and uncertainty without treating the AI interface itself as the source of the underlying numerical claim.

Exhibit: AI's role in the measurement architecture

AI function	Contribution	Evidentiary boundary
Integration	Connects and organizes evidence across methods, datasets and update cycles.	Does not make unlike claims equivalent or dependent evidence independent.
Criticism	Automates diagnostics, sensitivity testing and identification of contradictions or unstable assumptions.	Does not determine whether the underlying causal design is valid.
Adaptation	Updates models as new experiments and outcomes arrive and identifies where additional evidence would be valuable.	Model updates inherit causal credibility only from the evidence supporting them.
Decision support	Makes complex evidence and uncertainty easier to interrogate and apply to decisions.	The interface does not become the source of the numerical or causal claim.

These capabilities can materially improve how quickly and systematically the measurement system learns. None changes the standard required for causal identification.

Making creative a measurable input

Creative should enter the evidence architecture as part of the advertising treatment rather than remain an unobserved source of variance. Persistent creative identifiers, asset-level linkage and structured metadata can connect executions with delivery records, while computer-vision, audio, language and multimodal models can describe message, branding, imagery, sound, format and other executorial characteristics at scale.

This information can improve attribution, MMM, experiments and brand measurement by making material changes in the stimulus visible. It can help distinguish apparent publisher, audience or period effects from differences in the advertising delivered and generate hypotheses about which creative characteristics merit direct testing.

The evidentiary boundary remains important. Measured and model-predicted creative responses are different forms of evidence, and association between a creative characteristic and stronger outcomes does not establish that changing that characteristic would cause the predicted improvement. Creative data should make the treatment more explicit; causal creative claims still require a design capable of identifying the effect.

Creative data should make the treatment more explicit; causal creative claims still require a design capable of identifying the effect.

Integration: making the evidence system operable

The first role for AI is integration. Attribution, experiments, MMM, brand research and customer evidence are generated at different levels, on different schedules and with different definitions and assumptions. AI can help organize these inputs, identify differences in scope and make the relationships among them easier to interrogate.

This is principally an operational advantage. A system can connect campaign-level experimental evidence with granular attribution outputs, compare an MMM estimate with recent platform tests, flag a change in identity coverage or creative treatment, and surface where two apparently conflicting estimates actually refer to different populations, outcomes or periods.

AI can also help make complex measurement systems easier to use. Large language models can provide an interface through which decision-makers ask questions of validated statistical models and underlying evidence. But the distinction between interface and inference should remain explicit: the language model can help formulate, execute and explain an analytical query; it should not become the source of the numerical result or the evidentiary claim.

Where natural-language interfaces are used, the architecture should preserve a clear separation between the model that generates the evidence, the controls governing how that model can be interrogated and the interface used to communicate the result.

This distinction suggests three layers:

1. **A statistical layer**, containing the data, assumptions, model code and computations.
2. **A control layer**, determining which operations the AI can initiate and recording changes to specifications.
3. **A communication layer**, using language models to explain results, sensitivities and limitations in accessible terms.

The statistical model, rather than the language model, should remain the system of record. Numerical claims should be traceable to a defined dataset, specification and model run.

Criticism: automating challenge, not merely production

AI may be especially valuable in **model criticism**. Measurement models contain large numbers of assumptions that are difficult to test continuously by hand. An AI-enabled measurement system could repeatedly examine whether parameters are stable, whether channels are separately identified, whether conclusions depend heavily on particular priors or controls, whether response curves are plausible, whether an observational estimate predicts subsequent experimental results and whether changes in identity coverage or data collection have materially altered the evidence base.

The important shift is from automated production of estimates to automated challenge of estimates. A system might report, for example, that a channel's estimated effect is positive under most specifications but highly dependent on one control variable and inconsistent with two recent experiments. That is more informative than simply returning the preferred model's ROAS.

Automated criticism should not become automated approval. A model can pass technical diagnostics and still answer the wrong business question. A break in a statistical relationship may represent either a modeling failure or a genuine change in the marketplace. Human judgment remains essential, but AI can concentrate that judgment on the assumptions and conflicts that matter most.

The same role can extend to causal and quasi-experimental analysis. AI can help identify potential confounders, screen comparison markets, run placebo and sensitivity tests and surface alternative causal explanations. These are aids to research design and criticism rather than substitutes for identification. Whether an untreated counterfactual is credible remains a question of design, evidence and substantive judgment, not computational sophistication.

Adaptation: learning where advertising effects differ

A continuously learning measurement system should also improve its ability to identify where advertising effects differ across populations and conditions. Heterogeneous treatment-effect methods can help estimate differences across customer states, category participation, geography, creative exposure and other pre-treatment characteristics when they are supported by randomized or otherwise credibly identified evidence.

A conventional targeting model asks:

Who is most likely to buy?

A causal targeting model asks:

Whose probability of buying is most likely to change because they receive the advertising?

Those populations need not be the same. A loyal customer can have high purchase propensity and low incremental responsiveness, while a lapsed or occasional customer can have a lower baseline probability of purchase but more scope for advertising to change behavior.

Methods including causal forests and Bayesian hierarchical models can help estimate this variation, but their value comes from the causal evidence on which they are trained or constrained, not from model sophistication alone. Searching observational data for audience cells with high attributed returns can produce more granular bias rather than more granular truth.

The role of AI is to make this learning process easier to operate at scale: updating heterogeneous-effect estimates as new experiments arrive, identifying where predictions are unstable and helping determine which additional tests would be most informative.

Adaptation: experiment-trained attribution

Perhaps the clearest example of how the proposed architecture could evolve is **experiment-trained attribution**.

Observational attribution and experiments possess complementary strengths. Attribution is scalable, fast and granular but can be biased. Randomized experiments provide more credible estimates of incrementality and are typically estimated at broader levels. The emerging opportunity is to use experimental evidence to train and discipline scalable attribution models rather than treating observational attribution as self-validating.

Amazon's proposed multi-touch attribution methodology provides one illustration. Its researchers distinguish between machine-learning models trained on observational signals, which can generate precise and granular allocations but may not estimate incremental effects, and randomized trials, which estimate incrementality more credibly but generally at broader levels. The methodology combines the two so that experimentally estimated effects inform attribution credit allocation.

The conceptual shift is more important than any particular implementation. Providers of scalable attribution models should demonstrate that their estimates can predict experimental evidence that was not used during model estimation. Advertisers and their agencies should, in turn, treat this ability to generalize to independent experimental evidence as an important test of attribution-system credibility.

A mature system could operate as a continuous learning loop:

1. Randomized or strong quasi-experiments estimate incremental effects for selected campaigns, audiences or markets.
2. Models learn how those effects vary with campaign, audience, exposure, creative and contextual characteristics.
3. The models predict incremental effects in circumstances where direct experimentation is unavailable.
4. New experiments test those predictions.
5. Prediction errors update the model and identify where additional experiments would reduce the most decision-relevant uncertainty.

This turns experimentation from occasional validation into **measurement infrastructure**. A learning system can go further by using prediction error and remaining decision uncertainty to identify where the next experiment would be most valuable. AI can help automate that prioritization, but the experimental evidence remains the source of causal learning.

But experiment-trained attribution still has boundaries. Evidence from retail media may not generalize to national television. Lower-funnel experiments may not calibrate long-term brand effects. A platform experiment may observe only outcomes visible within that ecosystem. Training domains, estimands and limits to external validity must remain explicit.

The future is not experimentation once followed by automation forever. It is continuous interaction between experimentation and modeling.

This changes the meaning of speed in measurement. Daily data are valuable when they accelerate learning and improve operational decisions. They should not acquire greater causal authority merely because they arrive sooner. A real-time attribution metric should not overturn a strategic conclusion supported by stronger experimental and business-level evidence simply because the latter is updated less frequently.

As measurement becomes more tightly embedded in planning and in-flight optimization, decision speed becomes part of the measurement problem. Different forms of evidence update at different rates and carry different levels of causal authority. A measurement system needs explicit rules for which signals can trigger immediate tactical action and which decisions require stronger corroboration, so that faster updating does not automatically translate into faster propagation of measurement error.

AI can substantially reduce the operating cost and time required for this learning cycle. The objective should nevertheless remain faster accumulation of credible evidence, not simply faster production of numbers.

AI can amplify false precision

AI introduces a distinctive risk because it can make incomplete or uncertain evidence appear unusually coherent. Four forms of false precision are particularly important.

- **Data integration can be mistaken for causal completeness.** Combining large volumes of exposure, identity and transaction data does not create a counterfactual or recover important influences that were never observed.
- **Dependent evidence can appear independently corroborative.** Several estimates may look mutually reinforcing while depending on the same exposure logs, identity spine, conversion feed or model assumptions.
- **Granularity and statistical precision can outrun identification.** A campaign-level causal result can be distributed across audiences, creatives or touchpoints with considerable numerical precision even where the finer allocations remain model-dependent.
- **Explanation can be mistaken for verification.** An AI system can produce a persuasive account of why an estimate appears reasonable without providing any evidence that it survives an experiment, falsification test or credible alternative specification.

The appropriate response is not to reject AI but to design systems that make uncertainty harder to hide.

Criticism also requires independent evidence

The same discipline applies to the wider evidence architecture. AI can help compare and reconcile results, but it cannot make dependent estimates independent. The language of triangulation can itself produce false reassurance. Evidence should be weighted according to its design quality, relevance and independence, not simply counted.

For every major estimate, measurement providers – including platforms where they report their own effectiveness estimates – should make its evidential basis visible to advertisers and their measurement partners. At a minimum, this should include:

- The underlying exposure and outcome data.
- The included and excluded population.
- The identity and matching methodology.
- The comparison or counterfactual.
- The unit of analysis.
- The time horizon.
- The principal modeling assumptions.
- Whether the evidence is experimental or observational.
- Which other estimates depend upon the same inputs.

Advertisers and agencies should use these disclosures to determine not only the apparent strength of an estimate, but also the extent to which supposedly separate estimates actually provide independent evidence.

A well-powered randomized experiment may provide more information than several observational analyses derived from the same platform dataset. Equally, a narrow platform experiment may provide less useful evidence for a national, long-term budget decision than a well-specified MMM supported by external validation.

Triangulation is not the process of averaging competing answers. It is the process of understanding why different forms of evidence do or do not agree.

Decision support: different questions require different evidence

The architecture should not be implemented identically for every advertiser or consumer journey. For a direct-response subscription service with a short and highly observable digital path, granular attribution may provide considerable operational value, supported by frequent conversion-lift experiments and a relatively fast MMM.

For a frequently purchased consumer good, transaction history and retailer data may become central. Household or geographic experiments may be particularly important for distinguishing existing brand loyalty from incremental advertising response.

For an automobile, financial product or other long-consideration purchase, exposure-to-sale attribution may be much less informative. Brand measures, search, inquiries, intermediate behaviour, geographic variation and longer-horizon modeling assume greater importance.

The aim is not methodological uniformity. It is consistency in the standards by which evidence is judged.

Advertisers and their agencies or measurement teams should begin each measurement program by specifying:

1. The decision to be made.
2. The outcome and counterfactual of interest.
3. The relevant population and time horizon.
4. The parts of the consumer journey that can and cannot be observed.
5. The principal threats to inference.
6. The combination of evidence most capable of reducing those threats.

AI can make this decision framework easier to use by helping teams interrogate scenarios, surface assumptions and identify which uncertainties could actually change the decision. It should not determine the evidentiary standard independently of the advertiser, agency or analyst responsible for the decision.

Starting with the advertiser's decision in this way helps prevent whichever dataset, platform or vendor product happens to be available from defining the measurement question.

Governance is part of the architecture

Automated measurement systems require governance as well as data and modeling. AI governance should sit inside measurement governance rather than alongside it as a separate regime: the same requirements for provenance, reproducibility, evidentiary scope, decision rights and accountability continue to apply when analytical tasks are automated.

Governance becomes more important, not less, as the architecture becomes more automated. Organizational adoption is not simply a question of analytical literacy; it also depends on decision rights, workflow design and clarity about who is authorized to act on different forms of evidence. Responsibility should be explicit. Advertisers should establish the business questions, evidential standards and decision rules that govern the system, working with their agencies and internal analytics teams where appropriate. Measurement providers should document methodologies, assumptions, data dependencies and model changes sufficiently for their estimates to be scrutinized and reproduced. Platforms, publishers, identity providers and other data owners should provide transparent information about the coverage, construction and limitations of the data on which measurement depends.

Across the program, these parties should work to establish common definitions of exposure, conversion, attribution, incrementality and causation; consistent rules for comparing methods; transparent reporting of identity coverage and match quality; experimental and model-validation standards; version control for data, models and assumptions; and a defined process for investigating contradictory results.

AI adds additional requirements. Providers and measurement teams deploying AI-enabled systems should log material model changes, automated decisions and data revisions; ensure that numerical outputs can be independently reproduced; and connect causal claims to identifiable comparison strategies rather than to the fluency of an AI-generated explanation. Human judgment should remain concentrated at the points where it matters most: defining the question, choosing the estimand, assessing assumptions, interpreting conflicts and determining what evidence is sufficient for action.

Commercial independence also matters. Platforms and publishers that both sell media and measure its effectiveness should provide sufficient methodological transparency and access to relevant outputs for their claims to be evaluated beyond their own systems. Advertisers, for their part, should seek independent validation where the decision warrants it and avoid allowing one seller to define both the intervention and the standard by which it is judged. Platform evidence can make an important contribution, but its commercial provenance should remain visible when evidence is weighed.

Privacy governance is equally integral. Advertisers, platforms, data providers and measurement companies should ensure that richer analytical capability does not come at the expense of appropriate privacy, security and data-governance safeguards. First-party data, clean rooms and probabilistic techniques can reduce the need to exchange unrestricted person-level information, but technical controls do not remove questions of lawful use, purpose limitation, data minimization, security and consumer expectations. Privacy, security, lawful use, purpose limitation and data minimization should be treated as design requirements within the measurement architecture rather than as downstream controls.

Integrated measurement also requires explicit analytical provenance. Users should be able to identify the principal data, models and assumptions underlying a decision-relevant estimate; understand which apparently separate results depend on the same exposure logs, identity graph, conversion data or prior assumptions; and distinguish genuinely independent corroboration from repeated analysis of the same underlying evidence. Material changes in data sources, model specification or calibration should be documented so that conclusions can be reproduced, challenged and revalidated where necessary.

What next-generation measurement should deliver

The central promise of this architecture is not perfect attribution. Perfect reconstruction of every consumer journey is neither technically realistic nor necessary for most business decisions. The more useful objective is better-bounded inference: establishing whether an effect is likely to be incremental, the credible range of its magnitude, the circumstances in which it applies and whether remaining uncertainty could materially change the decision.

A mature system should produce four related forms of output:

1. **Descriptive truth:** What advertising was delivered, to whom and with what observable response.
2. **Causal evidence:** What changed because of defined advertising interventions.
3. **Business-level explanation:** How advertising fits alongside other drivers of demand over the short and long term.
4. **Decision guidance:** What action is justified given the evidence, uncertainty, costs and risks.

Point estimates will sometimes remain necessary for planning and forecasting. Measurement providers and analytics teams should accompany decision-relevant point estimates with appropriate ranges or scenarios, sensitivity analysis and a clear explanation of the assumptions that materially affect the conclusion. Advertisers and agencies should expect this information when comparing estimates and deciding whether the remaining uncertainty matters to the decision.

Uncertainty should be treated as an output of good measurement, not as evidence that measurement has failed.

Conclusion

The measurement-confidence gap cannot be closed by greater data granularity, stronger causal design or longer-term measurement in isolation. Each solves a different part of the problem. Greater observability expands what can be measured; experiments strengthen what can be inferred causally; brand, customer and aggregate evidence extend the outcome and time horizon; and attribution provides the operational resolution required for many day-to-day decisions. Next-generation outcomes measurement needs to integrate these forms of evidence rather than force any one of them to perform every role.

AI can make this architecture faster, less costly and more accessible by helping to integrate evidence, automate model criticism, update models as new information arrives and make uncertainty easier for decision-makers to interrogate. It does not constitute a fourth measurement paradigm alongside attribution, experimentation and MMM. Prediction is not incrementality, model sophistication is not causal identification, and greater computational capacity does not create a counterfactual.

Systems should be judged by whether assumptions are visible, dependent evidence is distinguished from independent corroboration, stronger causal evidence constrains weaker evidence and disagreements between methods are exposed rather than averaged away. The aim is not a single definitive ROAS, but a continuously updated evidence system that shows what can be concluded, at what level of confidence and what additional evidence would most improve the decision. The test for AI-enabled measurement should be whether it improves the discipline with which evidence is combined and challenged, not whether it produces a more seamless or apparently definitive answer.

This chapter has addressed the architecture of evidence: how different forms of measurement can observe, identify, contextualize, test and constrain one another. That does not by itself solve the marketplace problem. Even a strong evidence system can remain costly and difficult to operationalize if every seller requires different data connections, definitions, identity processes, reports and optimization interfaces. Chapter 10 shifts from the architecture of evidence to the architecture through which credible outcome signals are used across the television marketplace.

Chapter 10:

From outcomes capability to marketplace scale

Television already has substantial outcomes capabilities; the remaining challenge is cross-seller scale. The market already has substantial capabilities for attribution, incrementality, optimization and, increasingly, outcome guarantees. The constraint is that advertisers often have to recreate the connections, definitions, measurement specifications and workflows each time they move from one seller or environment to another.

Individual sellers increasingly support outcomes measurement, in-flight optimization and outcome-based commercial propositions, but advertisers may still need to reconnect data, redefine outcomes, reconcile measurement specifications and rebuild workflows across environments. The relevant question is what minimum interoperability would reduce that duplication without eliminating legitimate competition or methodological differentiation.

Outcomes-oriented measurement, optimization and commercial models are becoming more readily available: A+E Global Media has moved toward advertiser-specific outcome guarantees; AMC Networks has incorporated real-time outcomes attribution and optimization into its Audience+ platform; tvScientific allows qualified advertisers to establish a target cost per outcome and offers a model in which advertisers pay when the specified result is delivered; Universal Ads is building buying, measurement and performance optimization into a self-service platform spanning streaming and linear television²⁶.

The question is what would be required for television's existing outcomes capabilities to operate more consistently across multiple sellers while preserving legitimate differences in methodology, technology and commercial model.

Integration should not require television to reproduce the evidentiary logic of a closed platform. A multi-seller television marketplace has an opportunity to combine easier workflows with clearer distinctions between attributed and incremental outcomes, greater transparency around measurement assumptions and more systematic use of credible experimental evidence. In that model, causal confidence becomes part of the marketplace proposition rather than an obstacle to performance measurement.

This is a problem of scale, consistency and interoperability. Individual television companies have developed sophisticated outcomes capabilities, but advertisers buy across publishers, distributors, streaming services and platforms. Those environments can use different exposure data, identity systems, outcome partners, attribution rules, optimization technologies, reporting conventions and commercial terms.

The next phase of television outcomes depends less upon proving that individual solutions can work than upon **turning a collection of outcomes capabilities into a more coherent marketplace.**

²⁶ Sources: iSpot and A+E Global Media, *The Power of Outcome Guarantees for Advertisers* (March 2026); AMC Global Media, *AMC Networks Adds Key Verticals to Its Outcomes Attribution Product* (March 2026); tvScientific, *tvScientific Launches Guaranteed Outcomes* (June 2025); Universal Ads, *Comcast's Universal Ads Makes Linear TV Self-Service, Bringing Incremental TV Reach to Advertisers of Any Size* (April 2026).

From proof of concept to marketplace infrastructure

Many of these components are already operational. Multiple sellers and platforms now offer attribution, in-flight outcomes optimization or outcome-based guarantees, although these capabilities differ materially in scope, methodology and commercial structure. Publishers can increasingly identify audiences against advertiser objectives, measure business results, feed outcome signals into campaign delivery and optimize inventory while campaigns remain in market.

Advertiser demand is moving in the same direction. iSpot's 2026 video-advertising survey reported that more than 45% of marketers identified business outcomes as the most important factor in buying and negotiating media. Yet 47% also described measuring those outcomes across linear and streaming television as a significant challenge²⁷.

However, demand for outcomes appears to be developing faster than the consistency of the infrastructure required to support it. An advertiser running a campaign across several television publishers may be able to obtain business-outcome measurement from each of them. But that does not necessarily mean the advertiser can define an outcome once, connect its data once, compare results on a like-for-like basis, or allow a common outcome signal to inform optimization across the entire campaign.

This creates an important distinction between **capability** and **marketplace capability**. TV may collectively possess extensive outcomes capabilities while remaining more difficult to buy, measure and optimize collectively than a large digital platform.

Interoperability requires comparable meaning as well as connected systems

There are two related but distinct interoperability problems.

The first is **infrastructure interoperability**: whether advertiser data, exposure records, outcome signals and reporting can move efficiently across multiple television sellers without requiring a separate technical integration for each one.

The second is **measurement interoperability**: whether the resulting outputs are sufficiently transparent and comparable for advertisers to understand what each result represents and whether apparently similar measures can legitimately be compared.

The distinction matters because technical connectivity does not automatically create measurement equivalence. Two publishers could accept the same advertiser conversion signal through a common interface while using different definitions of advertising exposure, different identity universes, different attribution windows, different rules for eligible impressions or different methodologies for assigning credit. Both systems could be technically interoperable while producing performance estimates that measure materially different things.

The objective should not be to eliminate these methodological differences. Different publishers and measurement providers should remain free to develop competing approaches. But the differences need to be sufficiently visible and interpretable that an advertiser can determine whether two results are genuinely comparable, require adjustment before comparison or answer different questions altogether.

Infrastructure interoperability allows systems to connect. Measurement interoperability allows users to understand what the connected results mean.

²⁷ Source: iSpot, *The Short List: Key Takeaways from iSpot's 2026 Video Ad Strategy Survey* (May 2026).

Exhibit: Two dimensions of outcomes interoperability

	Core question	Examples
Infrastructure interoperability	Can the required data and signals move across systems?	1PD connectivity, exposure data, conversion signals, APIs, clean rooms, reporting schemas, optimization inputs.
Measurement interoperability	Can users determine what the resulting measures actually mean?	Outcome definitions, exposure units, identity methods, attribution rules, counterfactuals, causal status, measurement windows, uncertainty.

Connectivity without semantic consistency creates transferable ambiguity; semantic consistency without connectivity creates understandable but operationally isolated evidence.

A scalable television outcomes marketplace ultimately requires both. The industry is beginning to address the infrastructure side of this problem. In May 2026, nine major U.S. television publishers announced through OpenAP a shared initiative designed to connect advertiser first-party conversion data with exposure data across participating publishers. The proposed infrastructure includes a standardized conversion API, consistent exposure resolution and a common framework through which outcome signals can ultimately support attribution, optimization and automated buying across linear and streaming television²⁸.

The initiative represents a shift from publisher-specific outcomes capabilities toward shared outcomes infrastructure. It represents a move from **publisher-specific outcomes capability toward shared outcomes infrastructure**. Shared infrastructure can materially reduce the cost and complexity of connecting outcome data across publishers; the separate question is how comparable the measurement produced through those connections ultimately becomes.

Universal Ads has described the transition in similar terms: if the first phase of performance TV established that television can behave like a performance channel, the next phase requires greater consistency in outcomes, measurement and the confidence advertisers need to scale investment²⁹.

The relevant comparison is workflow integration

Major integrated platforms have helped establish advertiser expectations for advertising systems in which business objectives, audience selection, buying, conversion signals, measurement and optimization can operate through closely connected workflows. Google's automated bidding systems, for example, can optimize campaigns against conversions or advertiser-defined conversion values, while Amazon combines extensive first-party signals with media delivery, measurement, clean-room capabilities and automated optimization tools³⁰.

The important lesson is workflow integration rather than a simple comparison between television and digital. Where media delivery, identity, conversion signals, measurement and activation operate within the same technical environment, advertisers can often move more quickly from an outcome definition to measurement and optimization without rebuilding the process for each campaign component.

That integration is particularly powerful for behavior and transactions visible within the platform's own environment. Its scope should nevertheless be understood carefully. Outcomes occurring elsewhere – including activity across other publishers, retailers and offline environments – can introduce additional identity, coverage and causal challenges. Digital markets are themselves fragmented across platforms, devices, publishers, commerce environments and measurement boundaries.

28 Source: OpenAP, *National TV Publishers Partner to Establish Shared Infrastructure for Unified Outcomes and Audience Buying Through OpenAP* (May 2026).

29 Source: Universal Ads, *A Perspective on Performance TV: What 2025 Clarified – and What Comes Next* (March 2026).

30 Sources: Google Ads, *Value-Based Bidding Best Practices* (accessed August 2026); Amazon Ads, *Marketing Analytics: Measure Your Advertising Campaigns* (accessed August 2026).

The wider TV and video ecosystem has historically developed through a distributed structure. Inventory is spread across multiple sellers, while exposure data, advertiser first-party data, identity, transaction data and measurement may reside with different organizations. That diversity supports choice, specialization and independent measurement, but can also create repeated integration and reconciliation work for advertisers.

Greater consistency in how outcomes are defined, connected, measured and reported could improve cross-environment usability while allowing publishers and platforms to retain different data assets, technologies, methodologies and commercial models. Greater interoperability does not require a single architecture.

An outcomes layer, not necessarily a single new currency

The growing importance of outcomes does not require television to replace audience or impression-based transactions with a single universal outcomes currency.

Different measures perform different functions. Impressions can describe and price what was delivered. Audience measures can establish whom the advertising reached. Outcomes can influence valuation, measurement and optimization. Incrementality can help determine what advertising actually changed. In some circumstances, a business outcome can also become the basis of a commercial guarantee or payment.

A more useful way to think about the emerging marketplace is as a series of connected layers:

1. Impressions provide the delivery and transaction layer.
2. Outcomes provide an important measurement, valuation and optimization layer.
3. Incrementality provides a causal-validation layer.

Outcome guarantees and outcome-dependent payment can then sit on top of this architecture where sellers and advertisers choose to make those signals part of the commercial commitment.

This is more precise than describing every use of business results as an “outcomes currency”. A publisher that reports sales associated with television exposure is measuring outcomes. A seller that uses those sales signals to change campaign delivery is optimizing against outcomes. A guarantee introduces a financial commitment against a specified measure. A model in which payment itself depends upon acquisitions or sales moves further toward performance-based trading.

All are part of an outcomes-oriented marketplace. They need not represent the same commercial model.

The more important objective is to make these different approaches sufficiently **clear, consistent and interoperable** that advertisers can use them across the television marketplace.

What needs to become more consistent?

Scaling outcomes does not require every television company to sell the same product or adopt the same methodology. Publishers should continue to compete on content, audiences, inventory, data, technology, optimization capability, measurement partnerships and commercial models.

The requirement is narrower: sufficient interoperability at two levels. Some elements of the underlying infrastructure need to become easier to connect across sellers, while key measurement definitions and specifications need to be sufficiently transparent that advertisers can understand whether the resulting outputs are comparable. The objective is not identical outcomes measurement but connected systems and interpretable results.

The principal areas in which greater consistency or transparency could reduce marketplace friction are set out below.

Exhibit: The interoperability requirements of a scalable television outcomes marketplace

Area	What greater consistency would involve	Why it matters
Outcome definitions	Clear terminology and specifications for sales, visits, leads, installs, acquisitions, brand outcomes and other KPIs.	Apparently similar metrics need to be interpretable across providers.
Exposure definitions	Transparent and sufficiently comparable conventions for identifying advertising exposure across linear and streaming environments.	Every outcome analysis depends upon who is classified as exposed.
Identity and matching	Consistent disclosure of identity sources, matching methodology, coverage and known limitations.	Identity differences can change the population connecting exposure to outcomes.
Advertiser data connectivity	Standard interfaces through which advertiser first-party and conversion data can be made available across participating sellers.	Repeated one-to-one integrations increase cost and inhibit infrastructure interoperability.
Attribution specifications	Clear disclosure of attribution windows, eligible exposures, deduplication rules and credit-allocation methodology.	Different rules can produce materially different performance estimates even when the same underlying outcome data are used.
Incrementality	A consistent distinction between observed, attributed and incremental outcomes, with appropriate causal validation where incrementality is claimed.	Standardization should not imply that different levels of evidence are equivalent.
Reporting	More consistent schemas through which results can flow into advertiser, agency and measurement systems.	Outcomes become more useful when they can be compared and combined.
Optimization signals	Sufficiently standardized and timely outcome signals that can feed back into buying and delivery systems.	The value of outcomes increases when measurement can influence subsequent delivery.
Guarantee specifications	Clear description of what is guaranteed, how it is measured, principal exclusions and what happens if the commitment is not achieved.	Advertisers need to understand and compare different commercial propositions.
Governance and validation	Transparent methodologies and appropriate opportunities for independent verification.	Commercial dependence upon a metric raises the importance of trust and accountability.

This is fundamentally a two-part interoperability agenda. The first requirement is that systems can connect efficiently: advertiser data, exposure information, outcome signals and reporting should not require unnecessary bespoke integration across every seller. The second is that the outputs remain interpretable: differences in exposure, identity, attribution, evidence and scope should be visible enough for advertisers to understand whether results are comparable. Neither requirement implies identical measurement.

Standardize the interfaces, preserve competition and differentiation

The distinction between **infrastructure** and **proposition** is critical. A+E should be able to construct an outcome guarantee differently from AMC. tvScientific should be able to offer a cost-per-outcome model that differs from a traditional network guarantee. Publishers should be able to develop proprietary optimization technologies and differentiate on the performance of their inventory.

Greater consistency should not eliminate these differences. Instead, the industry should distinguish between what needs to be standardized and what needs to be made consistently interpretable. The plumbing may benefit from greater standardization: how advertiser outcome data enters the television ecosystem, how participating systems exchange signals and how results flow into reporting and optimization. Measurement methodologies need not be standardized in the same way, but their material specifications – including exposure definitions, identity and matching, attribution rules, outcome windows and evidentiary status – should be disclosed sufficiently consistently for advertisers to understand and compare the resulting claims.

The same principle already operates elsewhere in media markets. Common technical and measurement conventions can facilitate transactions without requiring sellers to offer identical inventory or commercial terms.

A mature TV outcomes marketplace might allow an advertiser to define a purchase, qualified lead, store visit or other business event once and make that signal securely available across several television sellers. Different publishers could then use the same underlying signal for different purposes: attribution, optimization, guarantees or future planning.

Consistency at the infrastructure level can support differentiation at the product level.

This provides a useful illustration of how different market structures can support outcomes measurement. Platform-controlled environments can offer powerful integration because exposure, identity and outcome signals can often be connected within one system. Cross-publisher and neutral architectures can extend measurement across participating organizations where compatible identity, permission, query and output arrangements can be established, although they require greater coordination and shared infrastructure.

A federated TV and video model could combine shared infrastructure with differentiated measurement, optimization and commercial propositions. Common technical connections need not require common products, methodologies or business models.

OpenAP's cross-publisher initiative is an early expression of this model. If advertiser conversion signals can enter through a standardized integration point and be translated into reusable outcomes signals across participating publishers, the advertiser does not need to recreate the same infrastructure separately for every seller³¹.

³¹ See: Open AP press release, National TV Publishers Partner to Establish Shared Infrastructure for Unified Outcomes and Audience Buying Through (OpenAP, May 13, 2026) at <https://www.openap.tv/press/cross-publisher-unified-outcomes-audience-buying>.

From individual guarantees to a federated outcomes marketplace

Outcome guarantees are already part of television's commercial proposition. The next question is how far guarantees can realistically operate across multiple sellers.

There is an important difference between four progressively more demanding propositions:

1. **Individual seller guarantees**, in which a network, platform or publisher guarantees a defined outcome against its own media.
2. **Comparable guarantees**, in which several sellers offer guarantees whose definitions and measurement specifications are sufficiently transparent for advertisers to understand and compare them.
3. **Multi-seller or pooled guarantees**, in which a defined group of participating sellers supports a common outcome commitment across a combined media package.
4. **A universal television outcomes guarantee**, in which one commercial outcome can be bought and guaranteed seamlessly across the wider linear, streaming and CTV ecosystem.

The first already exists. The second is principally an interoperability and disclosure problem. The third is possible in selected circumstances but requires materially greater technical, measurement and commercial coordination. The fourth would require a degree of alignment that is difficult to reconcile with the structure and diversity of the contemporary television marketplace.

Why a universal guarantee is so demanding

A universal guarantee is not simply an individual outcome guarantee repeated across more publishers. For a single commitment to operate across the wider television ecosystem, participating parties would need sufficiently compatible arrangements for at least:

- The business outcome being guaranteed.
- The qualifying population and transaction universe.
- What constitutes advertising exposure across linear and streaming environments.
- Identity and matching across devices, households, accounts and outcome records.
- Treatment of duplicated exposure across sellers.
- Attribution rules and measurement windows.
- Whether the guaranteed metric is observed, attributed or incremental.
- Treatment of creative, frequency and contextual differences.
- Data access and privacy.
- Reporting and reconciliation.
- Independent measurement or validation.
- Contractual responsibility.
- Financial remedies if the guarantee is missed.

If the commitment concerns **incremental** rather than attributed outcomes, the requirements become more demanding again because the commercial measure needs a credible counterfactual.

A universal television guarantee would require more than shared measurement infrastructure. It would approach a common commercial, measurement and settlement architecture across businesses that continue to differ in inventory, audiences, content, data, pricing, technology and measurement. Establishing such a model across the wider TV and video ecosystem would be exceptionally complex and would require a substantially greater degree of coordination than is needed to make outcomes measurement more consistent and interoperable across participating sellers.

The middle ground is more important

The alternative is not a choice between isolated publisher guarantees and one television-wide standard. A more plausible marketplace could combine shared infrastructure with selective multi-seller propositions. In such a model, participating publishers could use common or compatible mechanisms for:

Advertiser outcome signal

- permissioned data connection
- cross-publisher exposure resolution
- identity and matching
- agreed outcome measurement
- reporting and optimization

while continuing to differentiate through their inventory, content, audiences, pricing, optimization systems and commercial terms.

This can be understood as a **federated outcomes marketplace**: common infrastructure through which participating sellers can accept, resolve and measure appropriately permissioned advertiser outcome signals, combined with selective multi-seller commercial offerings built around inventory, outcomes and measurement arrangements that are sufficiently compatible.

The distinction matters. Federation does not require every publisher in the marketplace to participate, every outcome to be supported or every seller to use the same methodology. It creates a structure within which particular coalitions of sellers can offer interoperable or pooled propositions where the use case justifies the necessary coordination.

OpenAP's emerging cross-publisher infrastructure illustrates part of this direction. A standardized route for advertiser conversion signals, combined with exposure resolution across participating publishers, can reduce the need to rebuild the same basic data connection separately for each seller. That infrastructure could support attribution, optimization and, potentially, defined multi-seller commercial products. It does not by itself create a pooled guarantee or determine how commercial responsibility should be allocated among the participating publishers.

The hardest problem is responsibility

A pooled guarantee introduces a question that ordinary interoperability does not have to solve:

Who is responsible for the outcome?

Suppose a combined publisher package delivers 100 million impressions and is associated with 20,000 purchases. A simple allocation rule might assign outcome credit according to each publisher's share of impressions.

But **Publisher A = 30% of impressions does not necessarily imply Publisher A = 30% of incremental sales**

The foundational principles developed earlier in this paper explain why. Advertising response is heterogeneous. Frequency is nonlinear. Creative treatments can differ. Context and audience composition vary. Channels and publishers can reinforce one another. Some exposures may capture demand created elsewhere. And a campaign-level experiment can establish the incremental effect of the package without identifying the separate causal contribution of each component within it.

The attribution problem is a commercial-allocation problem as well as a measurement problem.

A pooled guarantee need not solve those two problems in the same way. For example, a credible experiment might establish:

Multi-seller campaign package → incremental sales

without establishing:

Publisher A = 35% of incremental sales
Publisher B = 25%
Publisher C = 22%
Publisher D = 18%

This does not make the package-level estimate deficient. It means the experimental design identified the effect of the package rather than the separate contribution of each participating seller. The publishers could still agree among themselves how commercial value or liability is allocated – through delivered media value, contracted share, agreed weighting, pricing or another settlement mechanism – without representing that allocation as a causal decomposition of the outcome.

This distinction could make multi-seller guarantees considerably more achievable. Commercial settlement does not have to masquerade as causal attribution.

Premium inventory should not become a homogeneous input

A second risk follows from the same principle. If pooled outcome systems allocate value only according to the number of impressions associated with near-term response, they can create pressure to treat inventory as a standardized performance input.

But an advertising exposure is not a homogeneous treatment. Audience, frequency, creative, context and timing can all change advertising response. Premium television inventory can also create effects through attention, memory, brand association, demand creation and longer-term value that may not be fully represented in an immediate conversion measure.

A multi-seller guarantee should not require participating publishers to surrender the contextual or commercial differentiation of their inventory.

One model could define an eligible portfolio of inventory and guarantee the performance of the portfolio while allowing individual sellers to retain differentiated pricing and commercial value.

The relevant proposition would be:

This defined premium-video package will deliver outcome X

rather than:

Every qualifying impression should be valued only according to its measured short-term conversion rate.

That distinction is important if outcomes trading is to complement rather than commoditize the wider value proposition of television.

Only some outcomes are suitable

The feasibility of a multi-seller guarantee will also depend heavily on the outcome and consumer journey. The strongest early candidates are likely to be outcomes that are sufficiently:

- Frequent.
- Observable.
- Timely.
- Matchable.
- Unambiguously defined.
- Economically meaningful.
- Numerous enough for stable measurement.

Depending on the advertiser and available data, these might include qualified website activity, leads, app activity, subscriptions, dealership or store visits, ticket purchases, selected retail transactions or other relatively short-cycle responses.

The challenge increases as outcomes become rarer, more delayed, less observable or more dependent on accumulated brand effects and non-media influences. A multi-seller guarantee for an observable short-cycle acquisition outcome is a fundamentally different proposition from guaranteeing long-term customer value, automobile purchases several months after exposure or brand-generated future demand.

Outcome guarantees should not become a universal replacement for audience or delivery guarantees. Their applicability should follow the consumer journey, business objective, measurement boundary and evidentiary requirements of the outcome being guaranteed.

Attributed and incremental guarantees are different products

The distinction between attribution and incrementality also needs to remain visible commercially.

A guarantee against:

10,000 attributed purchases

does not carry the same evidentiary burden as a guarantee against:

10,000 incremental purchases caused by the campaign.

The former principally requires an agreed method for connecting and assigning observed outcomes. The latter requires credible evidence about what would otherwise have happened.

Across a multi-seller package, an incremental guarantee may sometimes be supportable at the level of the package even when publisher-level incremental allocation is not.

This reinforces the central principle: **The level of commercial accountability should not imply a more granular causal claim than the evidence supports.**



AI can reduce coordination costs, not eliminate the identification problem

AI could make a federated outcomes marketplace substantially easier to operate. Automated systems could translate advertiser objectives into machine-readable outcome specifications, map data schemas across sellers, monitor identity and match coverage, reconcile reporting, identify inconsistencies in methodology, allocate spending dynamically within agreed campaign constraints and update models as new experimental evidence becomes available.

AI-enabled optimization could also move beyond a simple cheapest-cost-per-conversion objective. A decision system could optimize expected business outcomes subject to constraints around reach, frequency, premium inventory, context, audience coverage, creative and other advertiser requirements.

But AI does not resolve the central problem of responsibility. A model can estimate likely contribution among participating publishers; it cannot make those estimates causally identified merely by increasing the quantity of observational data or the sophistication of the algorithm. Experimentally anchored models may strengthen the allocation, but causal credibility remains bounded by the interventions for which credible causal variation exists.

AI can make federation more practical. It cannot make causal apportionment automatic.

The likely end-state is federation rather than universality

The strategic question is where greater coordination creates sufficient advertiser value to justify the additional complexity. The most plausible end-state is not a single television-wide outcomes market, but a layered one. Individual sellers would continue to offer their own outcome guarantees. Greater standardization would make those guarantees easier to understand and compare. Shared infrastructure could allow advertiser outcome signals to be used across multiple participating sellers. And, where the inventory, outcome definition, measurement approach and commercial arrangements are sufficiently compatible, groups of publishers could offer selective multi-seller guarantees.

These models could coexist. The market would not need every seller to support the same outcome, use the same methodology or participate in the same commercial structure. Instead, different levels of coordination could develop for different use cases, with stronger causal validation required where the commercial claim itself is incremental.

Exhibit: Four levels of outcomes-guarantee scale

Level	Commercial model	What is required	Principal difficulty
1. Individual seller guarantee	One seller commits against a defined outcome for its own inventory.	Clear outcome, exposure population, measurement methodology and remedy.	Evidentiary quality and transparency within the seller environment.
2. Comparable seller guarantees	Multiple sellers offer separately contracted guarantees that can be meaningfully compared.	Common disclosure fields and sufficiently interpretable definitions and specifications.	Similar metric labels may still represent different measurement populations or methodologies.
3. Federated / multi-seller guarantee	A defined group of sellers supports one outcome commitment across a combined package.	Compatible outcome, data, identity, exposure, measurement and contractual arrangements; agreed treatment of overlap and commercial responsibility.	Package-level performance does not automatically establish causal contribution by seller.
4. Universal television guarantee	One outcome commitment operates across the broad linear, streaming and CTV marketplace.	Extensive alignment across definitions, identity, measurement, data rights, governance, validation, settlement and commercial remedies.	Structural diversity, governance, incentives and the difficulty of causal and commercial apportionment make universal implementation highly demanding.

Shared infrastructure can support all four levels, but moving from technical interoperability to shared commercial liability introduces progressively greater measurement, governance and economic complexity.

A television marketplace organized in this way would be more interoperable without becoming homogeneous, and more outcomes-oriented without assuming that every campaign can be reduced to one universal performance measure.

The objective should be **interoperability without uniformity – and federation without universality**.

Consistency should not erase causality

Interoperability creates its own measurement risk. A signal can become easier to distribute, optimize and transact against without becoming more causal.

This distinction is fundamental. An outcome recorded after advertising exposure is not necessarily an outcome caused by the advertising. An attribution model may legitimately assign credit under a defined methodology without establishing how many additional conversions the campaign generated.

Likewise: The consumers most likely to convert are not necessarily those whose probability of conversion the advertising is most likely to change.

This issue matters even more when outcomes systems scale. If the same attributed-conversion signal feeds planning, optimization and guarantees across several publishers, systematic bias can also propagate through a larger part of the marketplace.

Not every outcomes system needs to produce an incremental estimate; observed, attributed and incremental outcomes answer different questions. Observed, attributed and incremental outcomes each answer useful but different questions.

The requirement should instead be consistent language about the level of evidence:

- An observed outcome records what happened after an exposure.
- An attributed outcome assigns some or all of the outcome to advertising under a defined model.
- An incremental outcome estimates how much additional behavior occurred relative to an appropriate counterfactual.

Those distinctions should remain visible as outcomes become more deeply embedded in television buying. There are already examples of architectures designed to connect attribution with causal evidence. Amazon's recent multi-touch attribution methodology combines machine-learning models with randomized experiments, using experimental evidence to help calibrate attribution rather than relying solely on observational associations. Television need not adopt one common methodology, but a scalable marketplace should make clear when similar-sounding outcomes claims rest upon different types of evidence.

Performance cannot mean only the nearest conversion

There is a second risk as television becomes more outcomes-oriented. The business signals easiest to incorporate into automated systems tend to be those that occur rapidly: site visits, leads, app installs, store visits and purchases.

These are valuable outcomes, and television should be able to compete directly for performance budgets tied to them, but television also creates effects earlier in the consumer journey and over longer time horizons.

If the outcomes marketplace concentrates only on immediately observable conversions, it can develop a structural bias toward demand capture. Optimization systems will naturally favor the behaviors included in their objective functions.

What is easiest to measure can gradually become what the market treats as most valuable.

This would be a mistake. Research reviewed elsewhere in this paper shows why short measurement windows can understate marketing's contribution. Google's *Effectiveness Equation*, for example, reports evidence that substantial advertising returns can emerge after the first several months, while wider analysis of attribution and MMM emphasizes the risk that attribution undervalues upper-funnel and longer-term effects.

A scalable outcomes marketplace should be capable of supporting both immediate performance objectives and effects that develop earlier in the journey or over longer periods.

Outcome systems should be capable of incorporating immediate response where appropriate while also connecting to brand effects, customer acquisition, customer value, incremental sales and longer-term business outcomes. No single measurement technique will accomplish this.

Attribution can provide rapid signals. Experiments can estimate the causal impact of defined interventions. MMM can place those effects within the wider marketing system and over longer periods. Brand research can capture effects that have not yet appeared in transaction data.

The strategic objective is not to make every campaign accountable to the same outcome.

It is to allow different advertiser objectives to connect consistently to the most appropriate evidence across the television marketplace.

From feasibility to scale

CIMM anticipated many elements of this agenda in its January 2026 draft study, *Developing a strategic roadmap for TV and video's outcomes future*. The proposed study included advertiser demand, mapping of the TV outcomes ecosystem, outcome predictability, optimization capabilities, the feasibility of guarantees at scale, structural barriers and the development of a strategic roadmap³².

The rapid development of the market suggests that the emphasis of such work should now evolve.

The central question is no longer whether an outcomes-focused television marketplace is feasible.

The market has demonstrated the feasibility of individual outcomes capabilities. The unresolved question is whether those capabilities can operate with materially less friction across multiple television sellers.

That work should address several connected questions:

- What is the minimum common infrastructure required for outcomes to operate efficiently across publishers?
- Which definitions, disclosures and interfaces need greater consistency, and which areas should remain proprietary and competitive?
- How can advertiser outcome signals be made securely reusable across multiple sellers rather than requiring repeated bilateral integration?
- How far can outcome guarantees realistically scale: from comparable individual-seller guarantees to selective pooled propositions across compatible sellers and what additional measurement, contractual and governance requirements would each step require?
- Can a multi-seller package be held accountable for a defined outcome without pretending that commercial settlement among participating publishers represents a causal decomposition of that outcome?
- How should attribution and optimization signals connect with experiments, MMM and brand measurement so that television does not develop a performance architecture optimized only toward the most immediate measurable response?
- What governance and independent-validation mechanisms would increase advertiser confidence sufficiently for outcomes-based television to scale?

These are no longer marginal questions about measurement innovation. They concern the architecture through which television competes for future advertising investment.

³² Source: CIMM, *Developing a strategic roadmap for TV and video's outcomes future* (January 2026).

From capability to a marketplace

Outcome measurement, optimization and guarantees are established in parts of the television market, and publishers and technology companies are beginning to connect these capabilities across organizational boundaries. The remaining question is how much consistency is required for advertisers to use them efficiently across sellers.

Greater consistency does not require a single measurement company, identity system, currency, optimization algorithm or commercial model. The relevant issue is whether common infrastructure in selected areas – data connections, definitions, disclosures, reporting and outcome signals – would reduce repeated integration and make different propositions easier to use and interpret.

Three forms of cross-industry work follow directly from this diagnosis: a set of common definitions and disclosure requirements for outcomes, including the additional information required for causal and incrementality claims; the development and validation of common practice guidelines for geographic randomized experiments in television and video, building on existing experimental standards and guidance; and a feasibility assessment of the interoperability required to scale television's existing outcomes capabilities across multiple sellers.

Chapter 11:

Priorities for practical cross-industry work on TV attribution and outcomes measurement

Two problems remain: outcomes claims are not always sufficiently transparent about what was measured, inferred and supported by the evidence, while credible outcomes capabilities can still be difficult to activate, interpret and compare across multiple television sellers. They require different remedies: interoperability cannot turn association into causation, while stronger causal evidence does not make fragmented workflows interoperable.

The analysis in the preceding chapters points to a clear next-stage agenda. Television and video do not need a new outcomes ecosystem to be created from scratch. They need targeted improvements that strengthen the credibility and interpretability of existing measurement while making credible outcome signals easier to use across the marketplace. The three proposals below form a practical roadmap for that next stage.

Three opportunities for industry collaboration follow:

1. The first is **establishing common definitions and disclosure requirements for outcomes measurement**: a common framework for describing what a material attribution or outcomes claim means, including an Incrementality Claim Specification that makes the causal quantity underlying any material incrementality claim explicit and consistently interpretable.
2. The second is **experimental practice**: developing and validating common practice guidelines for geographic randomized experiments in television and video, using existing standards and guidance as the methodological foundation.
3. The third is **marketplace interoperability**: assessing which definitions, interfaces, disclosures and workflows need greater consistency for credible outcome signals to operate efficiently across multiple television sellers.

The first would establish a common semantic layer for describing evidence and, for incrementality claims, a standard way of specifying the causal quantity actually estimated. The second would translate established causal and experimental principles into a common, operational practice framework for a defined and increasingly important television and video use case. The third would examine how interpretable outcome signals can move through the commercial and technical infrastructure of a multi-seller marketplace.

The three workstreams should remain proportionate and stage-gated, but the relevant gates differ. Greater coordination should not automatically imply a new accreditation regime, common methodology, shared infrastructure or governance body. Each initiative should build on existing work, demonstrate practical value and be tested before wider implementation.

The transparency framework can begin as a voluntary, demand-led set of common fields and definitions. The geo-RCT workstream should begin by mapping existing MRC, ARF, IAB, academic and practitioner requirements into a draft TV and video practice framework, identifying which provisions can be adopted by reference and where additional implementation guidance is required. The relevant test is not whether geo-RCT guidance should exist in principle, but whether the proposed framework adds sufficient practical value to improve the commissioning, execution and interpretation of real experiments. Draft requirements should be tested against completed and prospective studies before wider adoption. Marketplace interoperability requires a broader feasibility process involving advertisers, agencies, sellers, measurement companies, data providers and relevant technical initiatives. The practical objective

is a coordinated program: strengthen the evidentiary foundations, translate established principles into usable market practices where coordination can add value, and pursue wider standardization or shared infrastructure only where testing demonstrates benefits that exceed the implementation cost, governance risk and loss of useful differentiation.

Exhibit: Supporting TV's outcomes future – three areas for practical work

Workstream	First decision	Primary adoption mechanism	Principal implementation risk
1. Common definitions and disclosure requirements	What minimum information is required for a human or automated system to interpret a material outcomes claim correctly?	RFPs, briefs, contracts, methodology documentation, reporting and structured metadata.	Excessive burden, commercially sensitive disclosure or premature technical standardization.
2. TV/video Geo-RCT practice framework	Which existing requirements should be adopted by reference, and what additional TV/video-specific implementation requirements are needed to create a usable common practice framework?	Commissioning specifications, experimental-design and pre-specification templates, power-analysis requirements, treatment-integrity controls and common reporting fields.	Duplicating established guidance, over-prescribing legitimate methodological choices or encouraging experiments where geo-randomization is not fit for purpose.
3. Marketplace interoperability	Which sources of cross-seller friction create sufficient advertiser value to justify coordination?	Feasibility study, workflow pilots and staged implementation.	Misaligned incentives, data rights, governance complexity and unnecessary standardization.

Why coordination may be difficult

The principal barriers to a more mature outcomes marketplace are not purely technical. Participants occupy different positions in the commercial system and face different incentives.

Advertisers and agencies generally benefit from comparability, portability and transparency because those characteristics reduce switching costs and make cross-seller allocation easier. Media owners and platforms also benefit from credible outcome evidence but may have legitimate concerns that excessive standardization could eliminate methodological differentiation, expose commercially sensitive information or privilege measurement approaches better suited to some inventory than others. Measurement and data providers benefit from clearer requirements but may resist disclosures that reveal proprietary methods or make apparently differentiated outputs easier to compare. Identity providers and clean-room operators must reconcile interoperability with privacy, contractual and security restrictions.

The largest integrated platforms face a different structural calculus. Their advantage partly derives from operating media delivery, identity, first-party signals, conversion data, measurement and optimization within closely connected environments. Greater interoperability across independent television sellers could reduce some of that structural disadvantage, but recreating closed-platform integration across the open marketplace would introduce significant governance, privacy and competitive questions.

These incentives argue for minimum necessary interoperability rather than comprehensive standardization. The relevant test is whether a common element solves a demonstrable multi-party problem that participants cannot efficiently solve through bilateral arrangements, while leaving room for competition in methodology, technology, inventory and commercial models.

Proposal 1 – Measurement transparency: Establish common definitions and disclosure requirements for outcomes measurement

The case for a transparency framework does not depend on standardizing methodology. The marketplace spans different outcomes, consumer journeys, datasets, designs and use cases, and no single body can determine the appropriate method or confidence level for every claim.

The proposed framework would have a narrower purpose: establish the information that should accompany material attribution and outcomes claims so that users can understand what was measured, what was inferred, what evidence supports the inference and what material limitations remain.

Its purpose would be **disclosure rather than adjudication**. Different methodologies can legitimately answer different questions. The objective is to give advertisers, agencies and other users sufficient information to determine what a result represents and whether the underlying evidence is appropriate for the decision being made.

Build from existing standards rather than duplicate them

Any work in this area should begin with a detailed mapping of existing standards and guidance. In particular, the MRC's *Outcomes and Data Quality Standards* already establish substantial expectations covering attribution methodology, weighting and lookback windows, experimental design, MMM, model governance, data quality, matching, coverage, disclosure and auditability.

Even methodologically credible outcomes studies can produce results that are difficult to interpret or compare because the causal quantity, measurement boundary and evidentiary scope are not expressed consistently. The purpose of a transparency framework would be to identify the minimum information that advertisers and other users need in order to interpret material outcomes claims consistently across television and video services, determine which of those fields are already addressed adequately by existing standards, and isolate the genuine gaps created by contemporary TV and video workflows, cross-seller measurement and causal claims.

This would involve translating existing requirements into practical procurement and reporting fields and add a common semantic specification for material causal and incrementality claims. Where existing standards already prescribe the necessary methodological disclosure, those requirements should be incorporated by reference rather than rewritten.

Methodological quality and claim meaning are related but separate questions. A study can use a credible experimental or quasi-experimental design and still be described ambiguously if the intervention, comparison condition, outcome boundary or scope of the estimate is unclear. Conversely, transparent wording cannot rescue a weak counterfactual. Existing standards and accreditation processes principally address the quality of the underlying measurement; the proposed framework would make the meaning and evidentiary status of the resulting claim more explicit.

Candidate disclosure fields

The relevant disclosures would vary according to the outcome, methodology and use case. A framework could cover:

- **The claim:** What outcome is being reported, and is the result descriptive, attributed, modeled or intended as an estimate of incrementality?
- **Population and coverage:** Which audiences, households, devices, platforms or transactions are included, and which are outside the observable dataset? Which media exposures, publishers, devices or consumer groups cannot be observed?
- **Creative treatment:** Which creative executions were included, whether creative varied materially across the conditions being compared, and whether the reported result represents a media effect, a creative effect or the effect of a combined advertising treatment.
- **Exposure measurement:** How is advertising exposure established, and at what level – device, household, account or individual?
- **Creative assignment and optimization:** Was creative rotation fixed, randomized or determined by targeting or optimization systems?
- **Identity and matching:** How are exposure and outcome records connected? What proportion can be matched, and what assumptions are required in moving between devices, households and people?
- **Outcome measurement:** Is the outcome directly observed, inferred or modeled? What proportion of relevant transactions or behaviors is visible?
- **The counterfactual:** If a causal or incremental claim is being made, how is the estimate of what would have happened without the advertising constructed?
- **Identification strategy and evidentiary status:** If the result is presented as causal or incremental, what feature of the research design distinguishes the advertising effect from other influences on the outcome – for example random assignment, defined quasi-experimental variation, a model-based counterfactual or calibration to independent experimental evidence? If the result is observational, attributed or predictive rather than causally identified, is that status stated explicitly?
- **Prior behavior and selection:** What historical customer, purchase or engagement information is available, and how does the analysis address differences in pre-existing propensity between exposed and comparison populations?
- **Model dependence:** Which important assumptions, specifications or attribution rules materially influence the result?
- **Uncertainty and sensitivity:** Where relevant, what range of estimates is plausible, and how sensitive are the findings to alternative assumptions?
- **Generalizability:** To what population, campaign, platform or period can the result reasonably be generalized?
- **Analytical provenance:** Who supplied the data, conducted the analysis and funded the study, and was any independent validation undertaken?

Requirements should be proportionate to the claim. A QR-code response study does not require the same supporting information as a model estimating incremental automotive sales across multiple television and digital platforms. The relevant test is whether the information disclosed is material to understanding the result.

Additional specification for incrementality claims

Incrementality requires additional claim discipline because apparently similar results can estimate materially different causal quantities.

Every material incrementality claim should specify:

- **Intervention:** What advertising activity changed?
- **Reference condition:** What would the treatment be compared with – no focal advertising, lower spending, a different channel mix, a different targeting strategy or another defined alternative?
- **Population:** To which consumers, households, markets or other experimental population does the estimate apply?
- **Outcome boundary:** Which behavior or commercial outcome was measured, and which relevant outcomes or commerce environments were outside the observable data?
- **Surrounding media environment:** Which other material advertising activity remained present in treatment and reference conditions?
- **Treatment and observation period:** Over what period was advertising varied and for how long were subsequent effects observed?
- **Causal resolution:** At what level was causal variation actually created – campaign, platform, channel, market, household or another unit – and at what level is the incremental claim supported?

A practical reporting form could express the claim as:

Changing [defined advertising intervention] from [reference condition] to [treatment condition] was estimated to change [specified outcome] by [effect] for [defined population] over [specified period], while [relevant surrounding advertising activity] remained in place. The causal estimate is supported at [defined level].

This structure does not determine whether the underlying study is methodologically valid. It ensures that, once an incremental estimate has been produced, users can identify the causal quantity it represents and compare it only with estimates answering a sufficiently similar question.

A causal-claims module for incrementality

When a claim is intended to represent incrementality, the general disclosure framework should be supplemented by a small number of causal fields. At minimum, the claim should identify the intervention or change being evaluated; the counterfactual or reference condition; the identification strategy and any material identifying assumptions; the population to which the estimate applies; the outcome measured; the surrounding media environment that remained in place; and the treatment and observation period. The unit at which credible causal variation was created should constrain the granularity of the resulting claim.

This structure would distinguish materially different quantities that are currently described using similar language: total campaign incrementality, the incremental contribution of adding or removing a component, incremental ROAS, cross-environment or halo effects, post-treatment effects and incremental reach. Separate platform-level incrementality estimates should not automatically be treated as additive shares of a campaign-wide causal effect because the underlying populations, counterfactuals, media environments, outcome coverage and periods may differ.

A standardized claim structure could take the following form:

Changing [defined intervention] from [reference condition] to [treatment condition] was estimated to change [specified outcome] by [effect] for [defined population] over [specified period], while [relevant surrounding media activity] remained in place.

The objective would be semantic consistency rather than methodological uniformity. Different valid methodologies could continue to answer different questions, but the resulting claim would carry enough information for users to understand the causal quantity being reported.

Exhibit: Additional terminology for causal and incrementality claims

Term	Meaning
Identification strategy	The feature of the research design or analytical framework that allows the estimated effect of advertising to be distinguished from association with other causes of the outcome. Examples include random assignment, credible quasi-experimental variation or explicitly specified counterfactual assumptions. Predictive accuracy alone does not identify a causal effect.
Counterfactual or reference condition	The media or advertising condition against which the treatment is compared: for example, no advertising, existing spending, a lower spending level or the surrounding media mix without a particular component.
Incremental effect or lift	The estimated absolute or percentage change in a specified outcome caused by the intervention relative to the reference condition.
Incremental contribution of a component	The additional effect associated with adding, removing or changing a defined channel, platform, tactic or spending level within a specified surrounding media mix.
Total campaign or package incrementality	The incremental effect of a defined combination of advertising activity relative to a campaign-level counterfactual. A package-level result should not automatically be decomposed into component-level causal shares.
Incremental ROAS	Incremental revenue, profit or another specified economic outcome divided by the corresponding investment or change in investment.
Cross-environment or halo effect	A causally estimated effect appearing outside the environment directly promoted or measured. Observation of outcomes elsewhere is not, by itself, a halo effect.
Post-treatment or carryover effect	Incremental effects observed after active treatment ends. The post-treatment observation period should be stated explicitly.
Incremental reach	Additional people or households reached by adding a media component. This should be distinguished from incremental business outcomes.

Three additional disciplines are important. First, an incrementality estimate is conditional on its surrounding media environment: the incremental effect of CTV when added to linear television and search need not equal the effect of CTV under a different media mix. That difference reflects a different causal question rather than a defect in the experiment.

Second, the outcome boundary and time horizon form part of the estimand. A result based on direct-to-consumer transactions should not be interpreted as an omnichannel effect unless the relevant commerce environments are observed; similarly, a four-week and a twelve-week effect may represent different causal quantities rather than competing estimates of the same one.

Third, the causal claim should not become more granular than the causal variation supporting it. If an experiment randomizes an entire campaign, its strongest causal finding concerns that campaign. Subsequent analysis by publisher, creative, audience or frequency may be informative, but the randomization of the package does not automatically create causal identification for every component within it.

The framework should also be capable of structured, machine-readable implementation. This does not necessarily require a new transport protocol or technical standard – a more proportionate first step would be to define stable fields and controlled terminology that can be represented consistently in a methodology document, RFP response, reporting interface, API or automated measurement system. Designing the framework in this way would make it more useful as measurement feeds automated planning, optimization and AI-enabled decision systems.

A concise vocabulary for incrementality claims

A causal-claims module should encourage precise use of a limited number of terms:

1. **Counterfactual or reference condition.** The advertising or media condition against which the treatment is being compared. This could be no advertising, existing spending, a lower spending level, a different media strategy or the surrounding media mix without a particular platform.
2. **Incremental effect or lift.** The absolute or percentage change in a specified outcome estimated to have been caused by the intervention relative to the reference condition. Both the outcome and the reference condition should be explicit.
3. **Incremental contribution of a component.** The additional effect associated with adding, removing or changing a defined channel, service, platform, tactic or spending level relative to a specified surrounding media mix.
4. **Total campaign or package incrementality.** The incremental effect of a defined combination of advertising activity relative to a campaign-level counterfactual. A package-level result should not automatically be decomposed into component-level causal shares.
5. **Incremental ROAS.** Incremental revenue, profit or another specified economic outcome divided by the corresponding advertising investment or change in investment. Both the numerator and the spending denominator should be defined, particularly where a test concerns a marginal increase rather than the channel's entire budget.
6. **Cross-environment or halo effect.** An incremental outcome appearing outside the environment directly promoted or measured – for example, CTV advertising producing incremental retail or marketplace sales. "Halo effect" should describe a causally estimated cross-environment effect rather than simply the observation that purchases occurred elsewhere.
7. **Post-treatment or carryover effect.** Incremental effects observed after active advertising treatment ends. The observation period should be stated explicitly rather than assuming that the economic effect ends when the media flight ends.
8. **Incremental reach.** Additional people or households reached by adding a media component. This is an important media-delivery concept, but it should be distinguished explicitly from incremental business outcomes: additional reach does not itself establish additional sales or other causal effects.

The objective is not to invent a new language where established terminology already exists. It is to make the causal quantity represented by a claim explicit and to map the resulting framework onto relevant existing MRC and other industry terminology wherever possible.

Incremental contribution depends on the surrounding media mix

In a fragmented television and video environment, channel contribution is conditional on the advertising that remains around it. Consumers encounter multiple platforms and channels, which may reinforce one another, substitute for one another or capture demand created elsewhere.

Consider an advertiser running linear television, CTV and search. Suppose an experiment randomly suppresses CTV for part of the eligible population while linear television and search continue.

Conceptually, the experiment estimates:

$$\text{Incremental contribution of CTV} = \text{Outcome with linear TV} + \text{search} + \text{CTV} \\ - \text{Outcome with linear TV} + \text{search}$$

This answers a specific and decision-relevant question: what additional value did CTV generate when added to this existing media environment?

It does not establish what CTV would have generated without linear television or search. Nor does it identify a fixed percentage of the total advertising effect that permanently “belongs” to CTV.

If television stimulates search, some of the additional purchases may eventually occur following branded search. If CTV and linear television reinforce one another, the effect of CTV when linear television is present may differ from its effect when linear television is absent. If removing one platform causes spend or consumer attention to move elsewhere, its measured contribution will depend partly on that substitution.

In other words:

Incremental effect of CTV given the existing media mix

need not equal:

Incremental effect of CTV in a different media mix

The difference reflects a different causal question, not a defect in the experiment.

Every incrementality claim should identify the change being tested

Statements such as:

“CTV generated 15% incremental sales.”

or:

“Platform A was 40% incremental.”

are underspecified. Neither statement provides enough information to determine what was actually established.

A fuller structure would be:

Changing [defined advertising intervention] from [reference condition] to [treatment condition] was estimated to change [specified outcome] by [effect] for [defined population] over [specified period], while [relevant surrounding media activity] remained in place.

These fields matter because “incrementality” is not a context-free property of a platform, channel or campaign. It is an estimate of the effect of a defined change relative to a defined alternative. The same platform can generate different, equally legitimate incremental estimates under different treatment conditions, populations, media mixes, outcomes or periods.

For example:

Adding Platform A CTV advertising to the advertiser’s existing linear television and digital campaign was estimated to increase total measured sales among the eligible population by X% during the six-week treatment and two-week post-treatment period, relative to a randomized holdout from Platform A. Linear television and the advertiser’s other national digital activity continued in both groups.

This wording makes clear that the experiment estimated the incremental contribution of Platform A on top of the surrounding campaign. It does not imply that Platform A generated the same share of the total campaign effect or would produce the same result under a different media mix.

Platform-specific incrementality estimates should not automatically be added together

Suppose an advertiser receives separate incrementality studies from three streaming platforms. Each experiment may be well designed within its own environment, and each may legitimately demonstrate an incremental effect.

That does not mean the three estimates can simply be added together to calculate the total incremental effect of streaming.

The experiments may concern different eligible populations. Different advertising may remain in place in the respective control groups. Outcome datasets may have different coverage. Measurement windows may differ. The platforms may also interact with one another and with the rest of the campaign.

Distinguishing **component incrementality** from **total campaign incrementality** helps prevent this error. Component experiments can provide causal evidence about specific decisions; broader experiments and models can provide evidence about the overall package. Separate component estimates should be treated as complementary evidence, not automatically as additive shares of a total effect. A common incrementality framework should treat additivity as a claim requiring evidence, not as a default property of separately reported lift estimates.

The outcome boundary should be part of the claim

An incrementality claim should identify where outcomes are observed. Advertising shown on a television screen may prompt a consumer to purchase through a retailer, marketplace, advertiser website or physical store. A platform study observing only one destination may estimate incrementality accurately within that destination while missing effects elsewhere.

Conversely, observing a purchase in another environment does not prove that the advertising caused it.

A result based only on direct-to-consumer sales should say so. An omnichannel study should identify which commerce environments are included. A claim about customer acquisition should distinguish new customers from repeat purchasing where the underlying data permit.

This matters particularly for television and upper-funnel video, where the eventual transaction may occur on a different device, platform or commerce environment from the advertising exposure.

Time is part of the causal claim

The measurement period is part of what is being estimated, not merely a reporting detail.

Advertising effects can emerge gradually and continue after active treatment ends. A four-week sales effect and a twelve-week sales effect may represent different causal quantities rather than competing estimates of the same one.

Studies should distinguish the active treatment period from any post-treatment observation period and state whether the result represents immediate, cumulative or longer-horizon effects.

No universal measurement window is appropriate. An app installation and an automobile purchase operate on different decision cycles. The treatment and post-treatment periods should be explicit.

The causal claim should not become more granular than the evidence

If an experiment randomizes an entire CTV campaign, its strongest causal finding concerns the CTV campaign that was randomized.

The resulting data may subsequently be segmented by publisher, creative, audience, frequency or other attributes. Those analyses may be informative. But the fact that the overall intervention was randomized does not mean that every component within it was independently randomized.

The level of the causal claim should follow the level at which credible causal variation was created. This guards against outputs that are more granular than the evidence supporting them.

Designed for human and machine interpretation

The same framework should be capable of structured, machine-readable implementation.

This would not require a new transport protocol or technical standard at the outset. A more proportionate first step would be to define stable fields and controlled terminology that can be represented consistently in a methodology document, RFP response, reporting interface, API or automated measurement system.

This becomes more important as attribution and outcomes evidence feeds automated planning, optimization and AI-enabled decision systems. A result should be capable of carrying relevant evidentiary context with it: the type of claim, population, outcome boundary, exposure unit, counterfactual, methodology, period and material limitations.

Machine-readability should be a design principle for the semantic framework rather than a separate standard-setting exercise.

Use in commercial practice

The framework would operate through normal commercial practice rather than certification or accreditation. Its value would depend on use in procurement, product documentation, research briefs, contracts, reporting and ongoing commercial relationships.

Application would vary by role:

- **Advertisers and agencies** could use relevant questions in RFPs, measurement briefs, supplier assessments, contracts and partner reviews.
- **Broadcasters, streaming services and other media owners** could use the principles when describing outcomes products, commissioned research and measurement capabilities.
- **Platforms and technology providers** could incorporate relevant fields into product documentation, methodology papers, APIs and customer reporting.
- **Measurement, data and analytics companies** could use the framework to explain data provenance, methodology, dependencies, limitations and the evidentiary basis for material claims.
- **Trade associations and research organizations** could use it in guidance, training and market education and encourage participants to apply the relevant principles in their own practices.

The framework should be designed primarily for incorporation into existing commercial processes – RFPs, measurement briefs, contracts, methodology documentation, supplier assessments and reporting – rather than as a separate accreditation process.

This approach would not require a central organization to police compliance. Nor would public endorsement imply that an organization had satisfied every element of the framework.

Organizations using it could publish a short implementation statement identifying:

- Which elements of the framework are relevant to their role or product.
- How those principles are reflected in their measurement practices or documentation.
- Any material limitations or exceptions.
- Where commercial partners can obtain further methodological information.

This would document how the framework is being used without creating a certification mark, register of approved providers or centralized conformity process.

A demand-led route to adoption

Commercial demand provides the most plausible route to adoption. If advertisers and agencies begin asking consistent questions about identity quality, outcome coverage, counterfactual construction, prior behavior, uncertainty, incrementality and generalizability, providers have a practical incentive to supply clearer answers. Greater voluntary disclosure by sellers and measurement companies could, in turn, create stronger expectations of comparable information elsewhere in the market.

The practical ambition should be that an advertiser requesting an “incrementality” result does not receive an unqualified metric label. The measurement brief should specify the causal question in advance, and the resulting report should return the corresponding claim fields with the estimate. If two providers report incremental ROAS against materially different counterfactuals, populations or outcome universes, the difference should be visible before the numbers are compared.

Supporting resources could include:

- Model RFP and due-diligence questions.
- Illustrative measurement-brief requirements.
- Role-specific disclosure checklists.
- Examples of proportionate methodology statements.
- A standard Incrementality Claim Specification and reporting template, including worked examples for campaign-, channel-, platform- and package-level claims.
- Guidance and worked examples for describing observed, attributed and incremental outcomes.
- Examples of appropriately bounded and overextended causal claims.
- Guidance on component versus campaign incrementality.
- Guidance on outcome boundaries and time horizons.
- Anonymized examples showing how the framework has been applied in practice.

These resources should build on existing technical standards, methodological guidance and commercial practices rather than attempting to replace them. The MRC’s accreditation model is also relevant here because accreditation applies to defined services, metrics and environments rather than operating as a blanket certification of an organization’s entire measurement portfolio. A provider can be accredited for a specific outcomes-related service or measurement scope without demonstrating conformance across every methodology addressed in the broader Outcomes and Data Quality Standards. This reinforces the case for a modular transparency framework that can be applied proportionately to the particular claims, methods and services being offered³³.

³³ Media Rating Council (MRC), Accreditation and Digital Accreditation listings, accessed September 2026. MRC accreditation is granted to defined measurement services and metrics within specified environments rather than to an organization as a whole. See also MRC, *Outcomes and Data Quality Standards*, Final, September 2022.

Design principles

The principal test should be whether the framework improves how outcomes measurement is bought, sold, documented and evaluated.

Its usefulness would depend on six characteristics:

1. **Practicality:** The framework should focus on information that materially helps users interpret measurement claims.
2. **Neutrality:** It should not favor particular methodologies, vendors, platforms or business models.
3. **Proportionality:** Disclosure expectations should reflect the nature and significance of the claim being made.
4. **Flexibility:** Different organizations should be able to implement the principles in ways appropriate to their roles and products.
5. **Compatibility:** The framework should build on existing standards and practices wherever possible rather than creating duplicative requirements.
6. **Structured implementation:** Core fields and terminology should be defined sufficiently consistently to support both human interpretation and, where useful, machine-readable implementation.

Implementation would face legitimate constraints. Some methodological details may be commercially sensitive. Smaller providers may have fewer resources to produce extensive documentation. Measurement companies may depend on data supplied by third parties and be unable to verify every upstream assumption.

The framework should distinguish between information an organization can reasonably disclose itself and dependencies that should instead be identified as limitations.

The appropriate threshold is sufficient disclosure for an informed user to understand the basis, scope and material limitations of the claim.

A practical path to adoption

An initial consultation could test the disclosure categories and causal-claim terminology with participants across the ecosystem and produce a first set of voluntary principles and supporting templates.

Adoption could then be assessed through use in RFPs, measurement briefs, methodology documentation, platform reporting, structured data interfaces and commercial discussions. A defined review after a period of market use could identify areas of ambiguity, unnecessary burden or limited applicability without requiring permanent governance machinery.

The purpose is straightforward: **Make enough of the underlying evidence visible for informed users and automated systems to understand the basis, scope and limitations of a material attribution or outcomes claim.**

Proposal 2 – Experimental practice: Develop and validate a common practice framework for geographic randomized experiments in television and video

Geographic randomized controlled trials can provide valuable causal evidence in television and cross-platform video, particularly where advertising effects need to be assessed across broad media environments and individual-level identity or exposure data are incomplete. Their credibility, however, depends heavily on experimental design, execution and interpretation.

The industry should consider developing a **common TV and video Geo-RCT Practice Framework**: a practical set of principles, minimum requirements and implementation tools for commissioning, designing, executing, analyzing and interpreting geographic randomized experiments.

This would not create a new theory of experimentation or another general incrementality standard. Much of the methodological foundation already exists. The MRC's Outcomes and Data Quality Standards address randomization, treatment and control definition, bias, sample bases, uncertainty, treatment validation and contamination. ARF's RCT21 initiative examined randomized testing across television and digital media. IAB and IAB Europe have addressed incrementality and geo-randomization, while academic and practitioner research provides substantial guidance on geographic assignment, statistical power, interference, paired designs and incremental return estimation.

The task should be synthesis, operationalization and validation rather than methodological reinvention.

Existing guidance should first be mapped into the proposed framework so that established requirements can be adopted or referenced rather than duplicated. That crosswalk should also identify where television and cross-platform video create implementation questions requiring more specific treatment.

The resulting framework should establish the minimum questions and requirements that need to be addressed before a geo-RCT can support a material causal claim. These should cover the causal question and estimand; the unit of randomization; geographic-unit selection and independence; randomization and treatment assignment; statistical power and minimum detectable effects; treatment definition and contrast; creative treatment; interference, spillover and contamination; concurrent media and commercial activity; outcome measurement and coverage; analysis of clustered data; pre-specification; uncertainty; and the permissible scope and generalizability of the conclusion.

Nor should the framework prescribe a single estimator, geographic unit, minimum number of markets or experimental design. Different campaigns, outcomes and media environments require different choices. The objective should instead be to establish the **minimum conditions, analytical disciplines and disclosures required for a geographic randomized experiment to support the causal claim being made.**

The purpose is to make rigorous geographic experimentation easier to specify, execute and evaluate consistently across the television and video marketplace, while preserving legitimate methodological flexibility.

What should the proposed Geo-RCT Practice Framework establish?

The framework should distinguish between different geographic designs rather than treating them as interchangeable. Randomized designs may include market-level, paired or clustered randomization where assignment is genuinely random. Other approaches, including matched-market and synthetic-control methods, construct comparison markets observationally and should be treated as quasi-experimental designs with different identification assumptions. Stepped or sequential designs require additional consideration of timing, carryover and changing market conditions.

The framework should establish the requirements that must be addressed when designing, executing and interpreting each type of geographic study, without prescribing a single preferred design.

1. Geographic-unit selection and independence

A geographic experimental unit should be sufficiently distinct to allow advertising treatment to differ meaningfully between treatment and reference areas while providing outcome data at an appropriate level.

Possible units can include local markets, states, metropolitan areas, counties, ZIP-code clusters or other geographically defined markets.

The appropriate choice will depend on the media being tested, how precisely it can be controlled geographically, the availability of outcome data and the extent of consumer movement or media spillover across boundaries.

Geographic units cannot simply be assumed to be independent. Movement, media distribution and other forms of interaction can create interference between them. The framework should establish criteria for selecting and justifying the experimental unit rather than nominate a preferred geography.

2. Feasibility and conditions for geographic randomization

Not every campaign can support a geographic RCT. The advertiser must be able to create a meaningful difference in advertising treatment between geographic units. Outcome measurement must operate sufficiently consistently across the relevant markets, and the campaign must be large enough for the intervention to produce a potentially detectable change in an aggregate outcome.

A geographic RCT may be inappropriate where advertising cannot be controlled geographically, important national activity overwhelms the treatment difference, outcomes are extremely rare or only a very small number of usable geographic units is available.

Randomized and non-randomized geographic studies should remain clearly distinguished. Matched-market and synthetic-control approaches can be useful when randomization is infeasible, but they rely on different identification assumptions and should not be described as RCTs merely because the unit of analysis is geographic.

3. Statistical power, minimum detectable effects and treatment intensity

Geo experiments often contain far fewer independent units than individual- or household-level RCTs, while markets can differ substantially in population, sales, media behavior and underlying outcome variance.

A universal minimum number of markets is unlikely to be useful. Feasibility should instead be assessed jointly through:

- The number and independence of available units.
- The expected treatment effect.
- Outcome variance.
- The degree of clustering.
- Treatment intensity.
- The minimum detectable effect.
- Required statistical power.

Simulation-based power analysis can be particularly useful where historical market-level outcome data are available. Proposed assignments and treatment levels can be simulated against historical patterns to estimate whether a commercially meaningful effect would be detectable under realistic levels of heterogeneity, seasonality and noise.

The analysis should specify the assumed effect size, minimum detectable effect, target power, historical period used and sensitivity to alternative assumptions.

Where the available markets or achievable treatment contrast cannot detect a commercially meaningful effect with adequate power, the proposed experiment is not fit for that question.

Treatment intensity should also be defined before launch. A no-advertising control, heavy-up design, alternative spending level or different media strategy represents a particular intervention.

A heavy-up experiment, for example, estimates the effect of the specified increase from one spending level to another. It does not establish the total effect of the entire channel or imply the same return at materially different investment levels.

4. Creative treatment and treatment consistency

Geographic experiments should identify whether creative execution is held constant, varies naturally as part of the campaign treatment or differs systematically between experimental conditions. Where creative varies between treatment and reference geographies, the resulting estimate may represent the effect of the combined advertising treatment – including both media and creative – rather than the effect of media delivery alone. That is not necessarily a design flaw, but it should be reflected in the estimand and in the language used to describe the result.

Where the objective is to estimate the causal effect of a particular creative execution, creative itself must be varied according to a design capable of supporting that inference. Observational differences in performance among creative executions should not inherit causal status simply because another component of the campaign was randomized.

5. Interference, contamination and concurrent activity

Geographic units are not automatically independent. Consumers cross market boundaries; broadcast and cable footprints do not always align with experimental geographies; digital targeting can leak across borders; and people may live, work and shop in different markets.

Other advertising and commercial activity also continues during most experiments. National television, search, social media, retail media, pricing, promotions and competitor activity may affect both treatment and control markets or introduce geographically uneven shocks.

The framework should address:

- Treatment leakage.
- Cross-market mobility.
- Overlapping media footprints.
- National-media contamination.
- Material concurrent advertising or commercial activity.

These risks should be documented before launch, monitored during the test and reflected in sensitivity analysis or interpretation, if they are material. Some interference can be prevented; some can be measured; some can only be acknowledged. The experiment should make those distinctions explicit.

6. Pre-specification and analytical commitments

Principal design choices should be specified before outcomes are observed.

Pre-specification should include, at minimum:

- The primary research question and estimand.
- The experimental population and geographic units.
- The randomization procedure.
- Treatment and reference conditions.
- The primary outcome.
- The pre-period, treatment period and any post-treatment observation period.
- The minimum detectable effect and target statistical power.
- The primary analytical method.
- Rules for exclusions or treatment non-compliance.
- Any secondary outcomes or subgroup analyses.
- Conditions under which an experiment may be stopped early.

The objective is to prevent analytical choices from being revised retrospectively in response to observed results.

Any sequential monitoring or early-stopping rules should also account for the increased risk of false conclusions created by repeated examination of interim results.

Guidance can require pre-specification while remaining neutral among valid analytical methods.

7. Analysis of clustered geographic outcomes

Geographic RCTs differ statistically from experiments involving very large numbers of independent individuals. Observations within a market may be correlated; market sizes can vary substantially; outcomes may exhibit heavy tails; and repeated observations from the same geography are not equivalent to additional independent experimental units.

Analysis and uncertainty estimates should reflect these properties. Depending on the design and available data, appropriate approaches may include paired geo estimators, robust non-parametric methods and other techniques explicitly designed to address market heterogeneity and relatively small numbers of independent experimental units.

The framework should not prescribe a single estimator. It should instead identify the statistical properties that the analysis must address and the information required to justify the chosen method.

8. Statistical uncertainty and interpretation

Randomization does not make an experimental estimate exact. Geo-RCT results should report effect sizes together with appropriate uncertainty estimates rather than reduce findings to a binary distinction between “significant” and “not significant.”

A statistically inconclusive result should also be distinguished from evidence of no effect. Where the experiment lacked sufficient power to detect commercially plausible effects, that limitation should be explicit.

Reporting should communicate the estimated effect, the uncertainty surrounding it and whether the experiment was sufficiently informative to distinguish commercially meaningful effects from statistical noise.

9. Scope of the causal claim and generalizability

A well-designed geographic RCT can provide strong causal evidence about the specific intervention that was randomized. The remaining question is how far that estimate can be interpreted or generalized beyond the treatment, geographic units, outcomes and period actually tested.

A geo experiment showing that a defined increase in television advertising caused additional sales across randomized markets during a particular period does not necessarily establish:

- The causal contribution of individual television impressions.
- Which households or consumers were persuaded.
- The separate contribution of every publisher or platform within the treatment.
- The effect of a different creative execution.
- The effect at substantially higher or lower spending levels.
- The effect in another season, category or competitive environment.
- The long-term effect beyond the period observed.
- The return that another advertiser should expect from the same medium.

These are limits on the **scope and generalizability of the estimate**, not necessarily weaknesses in causal identification.

The framework should require the causal claim to remain aligned with the unit of randomization, intervention, population, outcome and period actually tested. A campaign- or market-level randomization can support a strong causal conclusion about the intervention that was randomized without creating causal identification for every publisher, creative, audience or impression within it.

A practical output

The proposed work should result in a **TV and Video Geo-RCT Practice Framework and implementation toolkit** that can be incorporated directly into research and commercial workflows.

The framework should establish a core set of principles and minimum practice requirements for credible geographic randomized experiments in television and video, while building on existing MRC, ARF, IAB, academic and practitioner guidance rather than reproducing it.

Supporting resources should include:

- A crosswalk showing how the framework relates to existing MRC, ARF, IAB and other relevant guidance.
- A geo-RCT feasibility checklist for determining whether geographic randomization is appropriate and statistically viable for the question being asked.
- A standard experimental-design and pre-specification template covering the causal question, estimand, unit of randomization, treatment and reference conditions, treatment and observation periods, outcome definition and planned analysis.
- A simulation and power-analysis template covering statistical power, minimum detectable effects and treatment contrast.
- Guidance on geographic-unit selection, independence, pairing or stratification where appropriate.
- A treatment-integrity checklist covering delivered treatment contrast, non-compliance, leakage, spillover, contamination, creative variation and concurrent media or commercial activity.
- Minimum reporting fields covering experimental execution, outcome coverage, statistical uncertainty, analytical choices and material limits to interpretation or generalization.
- Guidance on pre-specification, stopping rules and the treatment of post-hoc analyses.
- Worked examples distinguishing genuinely randomized geo experiments from matched-market, synthetic-control and other quasi-experimental designs.
- Examples of appropriately bounded causal claims and claims that exceed what the experimental design has established.
- Model language that advertisers and agencies can incorporate into RFPs, measurement briefs and experimental commissioning specifications.

These outputs should standardize the questions, evidence and disclosures required for credible practice, not require every experiment to use the same estimator, geographic unit, market definition, treatment design or analytical method.

Validation in practice

The framework should then be tested against real television and video experiments. Draft requirements should first be applied retrospectively to completed studies to determine whether they identify meaningful differences in experimental quality and interpretation. Where practical, they should also be tested prospectively in new experiments, allowing the feasibility, design, power, treatment-integrity and reporting templates to be evaluated under operational conditions.

Following this validation, the framework should be refined through industry consultation and made available for voluntary adoption. Advertisers and agencies could incorporate its requirements into RFPs, measurement briefs and study specifications; media owners and platforms could use them when developing geo-testing capabilities; and measurement providers could use the common fields when documenting methods and reporting results.

The practical test is whether users can answer three questions:

1. Was the experiment capable of answering the causal question it was designed to address?
2. Was it designed, executed and analyzed in a way that preserves the validity of the randomized comparison?
3. Does the causal claim remain within the boundaries of the intervention, population, outcome, period and level of randomization actually tested?

The most practical route to adoption would be to embed these requirements in the workflows through which experiments are commissioned and used, so that the estimand and principal design choices are defined before results are produced rather than clarified after competing estimates appear.

After an initial period of market use, the framework should be reviewed to determine which elements are sufficiently useful and stable for wider adoption. Where appropriate, validated elements could subsequently be considered for incorporation into relevant industry standards or guidance.

The objective is a common operational basis for commissioning and evaluating television and video geo-RCTs: rigorous enough to improve the credibility and comparability of experimental evidence, but flexible enough to accommodate legitimate differences among campaigns, categories, outcomes and research designs.

Proposal 3 – Marketplace interoperability: Assess how to make the television outcomes marketplace more consistent, scalable and interoperable

Television already supports outcomes measurement, optimization and guarantees across multiple sellers. The unresolved issue is whether advertisers can use those capabilities collectively without repeatedly redefining outcomes, reconnecting data, reconciling identity and attribution approaches and rebuilding reporting or optimization workflows seller by seller.

The proposed workstream would test where selected forms of consistency – in definitions, disclosures, data interfaces and workflow conventions – would materially reduce advertiser and agency friction while preserving legitimate methodological, technical and commercial differentiation. The objective is not to create one outcome, provider, identity system or commercial model, but to determine the minimum interoperability required for television's existing outcomes capabilities to operate more efficiently at marketplace scale.

The central question that would be addressed is: **Which minimum common interfaces, definitions, disclosures and governance arrangements would materially reduce cross-seller friction without eliminating legitimate methodological, technological or commercial differentiation?**

CIMM's earlier work on television and video outcomes identified advertiser demand, ecosystem mapping, outcome predictability, optimization, guarantees at scale and structural barriers as important areas for investigation. The current market and evidence base now support a more specific feasibility assessment.

The market has demonstrated that outcomes measurement, optimization and guarantees can operate in television. The more useful question is what is required to make these capabilities easier to use collectively.

A stage-gated assessment would begin by mapping advertiser workflows, initiatives and sources of friction; identify a limited number of gaps with demonstrable advertiser or agency value and no adequate existing solution; and test those gaps through real multi-seller workflows. The first stage should map existing advertiser workflows, initiatives and sources of friction. The second should identify a limited number of interoperability gaps for which there is demonstrable advertiser or agency value and no adequate existing market solution. The third should test those gaps through real multi-seller advertiser workflows.

Only where the evidence demonstrates material benefit should the study proceed to recommendations for common interfaces, shared specifications, governance arrangements or further technical development.

At each stage, the test should be whether coordination solves a problem more efficiently than existing commercial solutions or competition can solve it independently.

What should the study examine?

The study should begin with the advertiser and agency workflow rather than with individual technologies. The relevant test is whether an advertiser can:

- Define a business objective and outcome → Make the necessary data available
- Activate media across multiple television sellers → Receive sufficiently interpretable measurement
- Use outcome signals in subsequent planning and optimization.

without rebuilding the process separately for each participant.

A feasibility assessment should examine at least seven areas.

1. Identify where fragmentation currently inhibits scale

The study should map the practical workflow for outcomes-based television across major categories of seller and inventory, including linear television, addressable television, streaming and CTV.

It should identify where advertisers and agencies currently require separate:

- Technical integrations.
- Outcome or exposure definitions.
- Data connections.
- Identity processes.
- Measurement specifications.
- Reporting workflows.
- Optimization interfaces.

The study should distinguish fragmentation that creates genuine competitive choice from fragmentation that creates avoidable cost and complexity.

The objective would not be to catalogue every outcomes product in the market. It would be to identify the specific points at which fragmentation prevents advertisers from using those capabilities efficiently across sellers.

2. Determine the minimum common outcomes infrastructure the market requires

The study should assess which elements genuinely require greater interoperability.

Potential areas include:

- Common or mappable outcome definitions.
- Transparent exposure definitions and units.
- Standard interfaces for advertiser first-party and conversion data.
- Identity and matching disclosures.
- Attribution specifications and measurement windows.
- Reporting schemas.
- Metadata describing whether results are observed, attributed or incremental.
- Mechanisms through which outcome signals can flow back into buying and optimization systems.
- Common fields describing outcome guarantees.

The assessment should distinguish among four different forms of coordination:

Standardize → Make interoperable → Disclose consistently → Leave competitive.

Not every source of variation requires standardization.

The proposed coordination would allow different methodologies and systems to operate together without requiring advertisers to reconstruct their measurement architecture for every transaction.

3. Establish what should remain competitive and differentiated

A common outcomes layer should not create a homogeneous television marketplace.

Publishers should remain able to differentiate through their:

- Content.
- Audiences.
- Inventory.
- First-party data.
- Identity assets.
- Optimization systems.
- Measurement partners.
- Creative capabilities.
- Commercial models.
- Ability to deliver advertiser results.

The feasibility study should determine where common infrastructure should stop.

A shared interface for receiving an advertiser outcome signal does not require every seller to use the same identity methodology. A common reporting schema does not require identical attribution models. Comparable disclosure around guarantees does not require sellers to offer the same commercial proposition.

The study should test whether proposed forms of consistency improve cross-seller usability without reducing useful innovation or creating an inappropriate single market convention.

4. Determine how advertiser outcome signals could become reusable across sellers

One of the strongest potential advantages of a common outcomes layer would be to reduce repeated bilateral data integration.

The study should examine the technical, privacy, contractual and governance requirements that would allow an advertiser to define an outcome – such as a purchase, qualified lead, store visit, subscription or other business event – and make an appropriately permissioned signal available for use across participating television sellers.

Different sellers might then use that signal differently: for measurement, optimization, audience construction, guarantees or future planning.

The assessment should examine whether such an architecture can operate without requiring unrestricted movement of person-level information and how clean rooms, secure APIs, identity services and other privacy-preserving approaches could contribute.

The objective would be reusability of the signal, not centralization of the underlying data.

5. Determine how far outcome guarantees can scale across sellers

Outcome guarantees already exist in the market. The first scaling requirement is comparability: advertisers should be able to understand what different guarantees represent and compare them without assuming that similar labels describe identical quantities.

A feasibility study should assess whether a common disclosure framework could identify, for each guarantee:

- The outcome being guaranteed.
- The unit in which performance is expressed.
- The eligible exposure population.
- The source and coverage of the outcome data.
- The identity and matching methodology.
- The relevant attribution or measurement window.
- Whether the metric is observed, attributed or incremental.
- Principal exclusions and assumptions.
- The measurement provider and validation arrangements.
- Treatment of material changes in campaign conditions.
- The commercial remedy if the commitment is not achieved.

But the study should go further and test the boundary between comparable individual guarantees and a genuine multi-seller guarantee.

It should examine whether a defined group of participating publishers could support one outcome commitment across a combined media package and, if so, what additional arrangements would be required for:

- Common or mappable outcome definitions.
- Eligible inventory.
- Exposure deduplication.
- Identity resolution.
- Common advertiser conversion signals.
- Measurement methodology.
- Overlapping exposure across participating sellers.
- Package-level incrementality where claimed.
- Reporting and reconciliation.
- Allocation of commercial responsibility.
- Remedies and settlement if the package misses its commitment.

The study should distinguish explicitly between **measurement of the package** and **allocation within the package**. A campaign-level experiment or other credible design might establish that a defined multi-seller package generated incremental outcomes without establishing the separate causal contribution of each publisher. In those circumstances, the participants could use an agreed commercial allocation mechanism without representing that settlement rule as causal attribution.

The feasibility assessment should ask: **Can a multi-seller outcomes proposition create useful package-level accountability without requiring causal decomposition at a level the available evidence cannot support?**

This distinction may materially change the feasibility of pooled guarantees.

The study should also test which outcomes and consumer journeys are realistic candidates. Frequent and rapidly observable outcomes may support pooled guarantees more readily than rare purchases, long-consideration decisions, accumulated brand effects or other outcomes whose measurement requires longer horizons and broader evidence.

The assessment would determine the circumstances in which shared infrastructure can support individual guarantees, comparable guarantees and selective multi-seller guarantees, and where broader coordination would add more complexity than commercial value:

individual guarantees → comparable guarantees → selective multi-seller guarantees

while identifying where the additional complexity of broader coordination exceeds the likely commercial value.

6. Determine how an interoperable outcomes layer should connect with incrementality and wider effectiveness measurement

Scaling outcomes infrastructure should preserve the evidentiary distinctions developed throughout this paper.

A common attributed-conversion signal can become easier to distribute and optimize against without becoming more causal. Indeed, if a biased outcome signal is incorporated into more buying and optimization systems, its consequences can become larger.

The feasibility study should examine how scalable attribution and optimization signals can connect with stronger forms of causal and longer-term evidence.

This should include:

- The role of experiments in validating or calibrating outcome signals.
- The relationship between platform- or publisher-level attribution and broader MMM.
- The use of brand and intermediate-effect measurement.
- Longer-term customer and business outcomes.
- The extent to which common metadata can identify whether a result represents an observed, attributed or incremental outcome.

The aim is preservation of the distinction between measures supported by materially different levels of evidence. It should be to prevent interoperability from creating a false equivalence between measures supported by materially different evidence.

7. Determine how the marketplace can support full-funnel outcomes rather than only immediate conversion

A scalable television outcomes marketplace should be capable of competing for direct-response and performance investment.

But it should not define performance solely through whichever behaviors can be observed most quickly.

The study should examine how the architecture can support different types of advertiser objective, including:

- Website visits and search.
- Leads and applications.
- App installs.
- Purchases.
- Customer acquisition.
- Retention and churn reduction.
- Customer value.
- Brand response and consideration.
- Longer-term demand and economic outcomes.

This is both a measurement and a competitive issue.

If television's outcomes infrastructure becomes optimized only around short-cycle conversions, it risks under-representing some of the effects through which television creates value and reproducing the same short-term biases identified elsewhere in the outcomes ecosystem.

A scalable TV outcomes marketplace should support multiple outcome horizons, with the measurement approach matched to the effect being claimed.

What would the study produce?

The practical output should be a **feasibility assessment and strategic roadmap**, rather than a new technical standard.

It could include:

- A map of the existing outcomes workflow and the principal points of fragmentation across the television marketplace.
- A definition of the minimum viable interoperability layer required to reduce that fragmentation.
- A distinction between capabilities that should be standardized, made interoperable, disclosed consistently or left fully competitive.
- A proposed common framework for describing outcomes, attribution specifications and individual and multi-seller guarantees, including the minimum additional requirements for pooled commercial accountability.
- An assessment of technical approaches for reusable advertiser outcome signals across multiple sellers.
- An assessment of whether selected publisher coalitions could support a federated multi-seller guarantee, including the measurement, contractual and settlement arrangements required and the outcomes for which such a model is realistically supportable.
- A review of relevant privacy, security, governance and data-permission requirements.
- A framework showing how attribution and optimization signals should connect with incrementality testing, MMM and longer-term effectiveness measurement.
- A small number of priority use cases against which the proposed architecture could be tested.
- At least one pilot designed specifically to test the distinction between package-level measurement and publisher-level allocation within a multi-seller outcome proposition.
- An assessment of the organizational, contractual and commercial barriers to adoption.
- A staged roadmap identifying where industry coordination could create material value and where existing initiatives already provide an appropriate solution.

The study should map existing market initiatives rather than assume that new infrastructure must be created.

Where shared APIs, data-collaboration environments, measurement standards or cross-publisher initiatives already address parts of the problem, the objective should be to assess their coverage, compatibility and potential for extension.

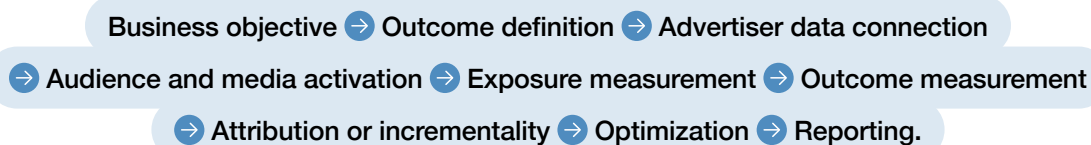
Test the proposition through real advertiser workflows

The assessment should be tested against real advertiser workflows. The study should test its conclusions against a small number of realistic advertiser use cases involving campaigns across several television and video environments.

For example, one pilot could examine a relatively frequent and rapidly observable outcome such as ecommerce purchasing or app acquisition. Another could examine a retailer or CPG outcome requiring data collaboration across publishers and commerce partners. A third could test a longer-funnel objective in which transactional outcomes need to be considered alongside brand or intermediate effects.

Where feasible, one pilot should also test a genuinely pooled proposition across a defined group of sellers: whether the participating publishers can accept one advertiser outcome signal, measure performance against one agreed package-level commitment and reconcile commercial responsibility without implying that the outcome has been causally decomposed across every publisher or impression.

For each case, the study could document:



The purpose would be to identify exactly where existing systems interoperate, where manual reconciliation or repeated integration remains necessary, and which changes would materially improve the advertiser experience.

This would also allow the study to distinguish between barriers that require industry coordination and barriers better resolved through competition and technological innovation.

A feasibility study, not a prescription

The feasibility study would assess whether selected components of existing outcomes infrastructure could operate more consistently across sellers, which components warrant coordination, what governance would be required and where existing market solutions are sufficient. It would not prescribe a single outcomes currency, identity system, attribution methodology, measurement provider or commercial model, or assume that all television campaigns should transact against outcomes.

That makes a feasibility study an appropriate next step. It would allow the industry to establish where greater interoperability offers material advertiser value before attempting to create standards or shared infrastructure. It could identify areas in which existing initiatives are already sufficient, areas in which greater alignment would be valuable and areas in which the cost or complexity of standardization would outweigh the benefit.

The test should be practical:

1. Would greater consistency materially reduce advertiser and agency friction?
2. Would it make television outcomes easier to compare, activate and optimize across sellers?
3. Could it be achieved without eliminating legitimate methodological and commercial differentiation?
4. Could the required data collaboration operate with appropriate privacy, security and governance?
5. Would the resulting outcome signals remain sufficiently transparent about their evidentiary status?
6. Would the architecture support both immediate performance outcomes and the wider range of effects television can create?
7. Would the resulting architecture materially improve cross-seller usability and make outcomes measurement easier to operate alongside the wider digital and cross-media environments in which advertisers already work?

Where the tests demonstrate material value, the study could advance a limited set of high-priority interoperability initiatives for market testing.

The objective is to determine what would be required for television's existing outcomes capabilities to operate as a more consistent, scalable and interoperable marketplace – and whether practical industry coordination can help achieve it.

TV's outcomes opportunity

The first two proposals strengthen the evidentiary basis of outcomes measurement by making claims more interpretable and geographic experiments more consistently designed and reported. The third addresses marketplace usability by examining where credible outcome signals can operate more efficiently across sellers. These are separate requirements. Interoperability does not strengthen the causal basis of a claim, while stronger causal evidence does not make a signal easier to deploy across sellers. Progress requires both clearer evidence and more usable infrastructure.

Conclusion:

From connection to confidence

The next frontier in outcomes measurement is not connection. It is confidence.

Television and video can connect advertising exposure with a wide range of consumer and commercial outcomes. This capability represents a major advance in television measurement and gives the marketplace a strong foundation on which to build. The harder question is what those connections establish.

A purchase after exposure is not necessarily a purchase caused by exposure. Better identity can show more confidently that two records belong together without showing that advertising changed the outcome. More granular attribution can describe the path to conversion in greater detail without producing equally granular causal knowledge. AI can integrate more evidence and interrogate it faster, but it cannot manufacture a counterfactual that the research design does not identify.

Claims need to be matched to appropriate evidence. Attribution can describe observable pathways and support operational decisions. Experiments can establish causal effects for defined interventions. MMM can place media within a broader cross-channel and longer-term business context. Brand research can capture effects that precede transactions. Customer and retailer evidence can show how purchasing, retention and value develop over time. Used together, these methods can form a stronger system of evidence, provided their differences remain visible and stronger causal designs receive greater weight when the claim itself is causal.

A scalable TV and video outcomes marketplace does not require one attribution model, dataset, outcomes currency or commercial model. It requires enough common language to make claims interpretable, enough experimental discipline to support incrementality claims and enough interoperability to make credible signals usable across sellers. What matters is sufficient common language to make material outcomes claims interpretable, sufficient experimental discipline to establish incrementality where it matters, and sufficient interoperability to make credible signals easier to use across sellers.

The same principle applies to outcome guarantees. A single market-wide guaranteed outcomes proposition across the full TV and video ecosystem would require extensive coordination over identity, attribution, measurement, reconciliation, liability and commercial responsibility. A federated model offers a more practical path: common or compatible infrastructure can make outcome signals and measurement easier to use across participating sellers, while individual companies retain their own commercial and technical approaches and groups of sellers develop joint propositions where doing so creates value.

Different market structures have different strengths. Integrated platforms can connect media, identity, conversion signals, measurement and optimization efficiently within their own environments. A distributed TV and video ecosystem can combine broad reach, varied inventory and multiple measurement approaches, but depends more heavily on cross-system consistency and coordination. The objective is not architectural uniformity, but evidence that is credible, interpretable and usable across the ecosystem.

Taken together, these disciplines provide a practical roadmap for the next stage of television outcomes measurement: preserve the innovation and diversity that have created today's capabilities, while making the resulting evidence more interpretable, more credible and easier to use across the marketplace.

The transition from connection to confidence depends on knowing what observed links mean, what remains uncertain and when the evidence is strong enough to support a decision.

Appendix:

Sources and References

The following sources are cited or substantively referenced in this Working Paper. Repeated citations in the text and footnotes have been consolidated here into a single reference for each source.

1. **Advertising Research Foundation (ARF)**, *RCT21: The ARF Initiative to Advance Cross-Platform ROI Analysis Through Application of Randomized Control Trial Measurement* (July 2020).
2. **Advertising Research Foundation (ARF) Cognition Council Cognitive Processes Working Group**, *The Brain on Advertising: Why We Notice, Feel, Remember and Act on Ads* (May 2026).
3. **AMC Networks**, *AMC Networks Adds Key Verticals to Its Outcomes Attribution Product* (March 2026).
4. **Amazon Ads**, *Amazon Marketing Cloud* (accessed August 2026).
5. **Amazon Ads**, *Marketing Analytics: Measure Your Advertising Campaigns* (accessed August 2026).
6. **Thomas Blake, Chris Nosko and Steven Tadelis**, *Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment* (January 2015).
7. **Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy and Steven L. Scott**, *Inferring Causal Impact Using Bayesian Structural Time-Series Models* (March 2015).
8. **Rick Bruner, Central Control**, *Hierarchy of Advertising Evidence* (2025).
9. **Central Control**, *How to Design a Geographic Randomized Controlled Trial: A Detailed Guide to “Rolling Thunder” and Other Transparent, Unbiased and Replicable Techniques for Measuring Incremental Advertising Impact with Experimental Design* (August 2025).
10. **Coalition for Innovative Media Measurement (CIMM)**, *Developing a Strategic Roadmap for TV and Video’s Outcomes Future* (January 2026).
11. **Michelle Dieschbourg**, *Data Clean Rooms Measurement: How Clean Rooms Solve Attribution After the Cookie* (July 2026).
12. **Ryan S. Elder, Gina Slejko, Jeff Dotson and Ariane Pol**, *Designing Delicious: An Examination of Creative Attributes Driving Food, Beverage, and Restaurant Advertising Effectiveness*, *Journal of Advertising Research*, Vol. 65, No. 1 (2025).
13. **Google**, *Ads Data Hub* (accessed August 2026).
14. **Google**, *The Effectiveness Equation* (2025).
15. **Google Ads**, *Value-Based Bidding Best Practices* (accessed August 2026).
16. **Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava and Dan Chapsky**, *A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook* (March 2019).
17. **Brett R. Gordon, Robert Moakler and Florian Zettelmeyer**, *Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement* (July 2023).
18. **Brett R. Gordon, Robert Moakler and Florian Zettelmeyer**, *Predicted Incrementality by Experimentation (PIE) for Ad Measurement*, NBER Working Paper No. 35044 (April 2026).
19. **IAB and IAB Europe**, *Guidelines for Incremental Measurement in Commerce Media* (November 2025).
20. **IAB Tech Lab**, *Data Clean Rooms: Guidance and Recommended Practices, Version 1.0* (July 2024).
21. **iSpot**, *The Short List: Key Takeaways from iSpot’s 2026 Video Ad Strategy Survey* (May 2026).

22. **iSpot and A+E Global Media**, *The Power of Outcome Guarantees for Advertisers* (March 2026).
23. **Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero and German Schnaidt**, *Amazon Ads Multi-Touch Attribution* (August 2025).
24. **Media Rating Council (MRC)**, *Outcomes and Data Quality Standards*, Final (September 2022).
25. **Media Rating Council (MRC)**, *Digital Platforms Accreditation Status* (July 9, 2026).
26. **Media Rating Council (MRC)**, *Accreditation and Digital Accreditation* listings (accessed September 2026).
27. **OpenAP**, *National TV Publishers Partner to Establish Shared Infrastructure for Unified Outcomes and Audience Buying Through OpenAP* (May 13, 2026).
28. **tvScientific**, *tvScientific Launches Guaranteed Outcomes* (June 2025).
29. **Universal Ads**, *A Perspective on Performance TV: What 2025 Clarified – and What Comes Next* (March 2026).
30. **Universal Ads**, *Comcast's Universal Ads Makes Linear TV Self-Service, Bringing Incremental TV Reach to Advertisers of Any Size* (April 2026).
31. **Jon Vaver and Jim Koehler**, *Measuring Ad Effectiveness Using Geo Experiments* (December 2011).
32. **WARC**, *What We Know About Attribution and Marketing Mix Modelling* (July 2026).
33. **Chris Williams and Jon Watts**, *Models as Masters? Marketing Mix Modeling and AI-Driven Media Decision Making* (July 2026).

cimm

Coalition for Innovative
Media Measurement

