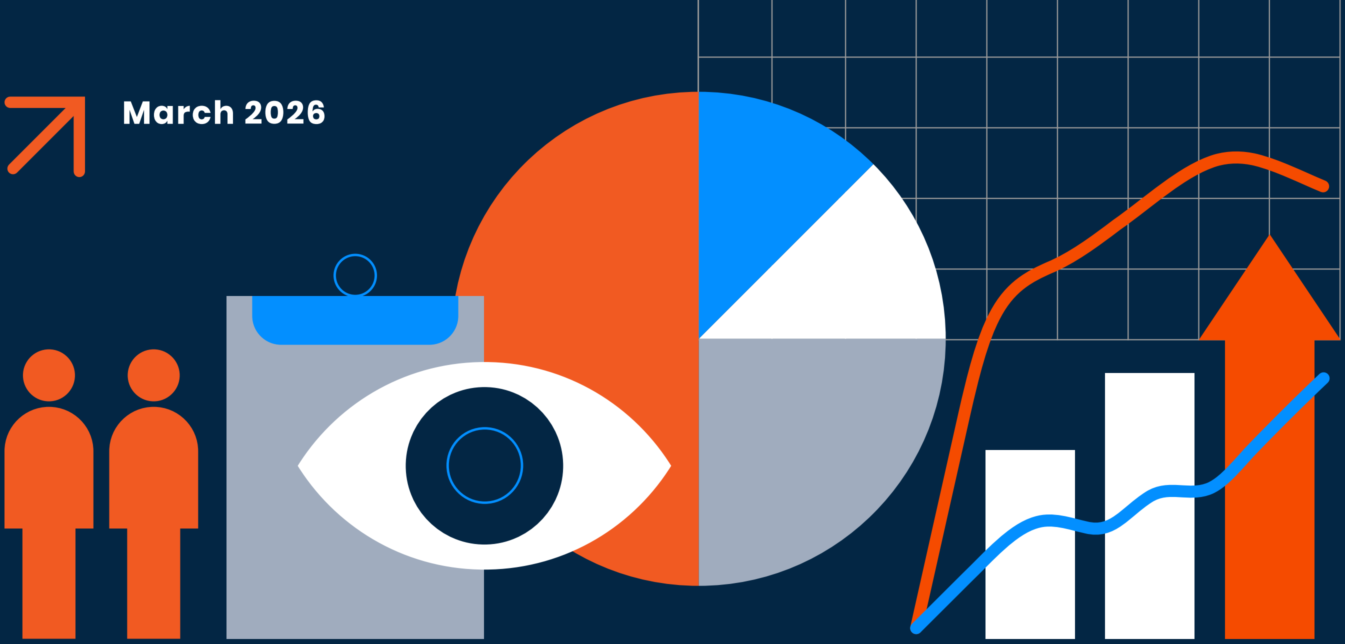
March 2026



Retail Media iROAS Demystified

Understanding Your Brand's *Incremental* ROAS

ovative

Albertsons Media COLLECTIVE

In partnership with professors from Northwestern University Kellogg School of Management.

marketing@ovative.com | mediacollective@albertsons.com

ovative.com
albertsonsmediacollective.com

Table of Contents

Context and Objectives

In our ROAS Demystified paper we unpacked the challenges of using Return on Ad Spend (ROAS) as a performance metric across retail media. We noted that incremental ROAS (iROAS) provides substantially more rigorous measurement of the business outcomes driven by media investment than standard ROAS, but it still lacks comparability across media networks and across channels within a single network.

While the definition of incrementality is simple, it can be applied and interpreted (and misapplied and misinterpreted) in numerous ways. In addition, there are a variety of ways to measure incremental impact. This creates confusion and pain points for advertisers and retail media Networks (RMNs) alike. To address this challenge, the objective of this paper is to:

1 Explain the major incrementality measurement approaches used across Retail Media Networks (RMNs)

2 Demonstrate how different methodological choices can materially change iROAS results using real-world data from 42 digital ad campaigns

3 Provide resources to advertisers and RMNs to drive better understanding of incrementality approaches

4 Create an industry-wide call to action to improve transparency in incrementality methodologies and build trust in the true business impact of retail media for advertisers

Part 1

What is Incrementality?

What is Incrementality?

The Interactive Advertising Bureau (IAB) defines incrementality as the measurement of the causal impact of marketing by identifying the additional business outcomes directly driven by a campaign or tactic, compared to what would have occurred in the absence of marketing activity. Simply put, incrementality isolates changes in customer behavior that occur *because* of media investment. If one group is exposed to an ad (test group) and another is not (control group), incrementality asks: to what degree do behaviors differ as a result of that exposure?

In retail media, the term incrementality is often used more broadly to describe impact across different time horizons and business contexts. In practice, incrementality may refer to the lift generated by a specific, time-bound campaign (e.g., a Back-to-School promotion), the ongoing incremental contribution of an always-on tactic (e.g., onsite search), or the incremental impact of reaching a particular audience segment (e.g., new-to-brand customers). While each of these use cases is valid, they are conceptually distinct and frequently measured using different assumptions, data scopes, and methodologies.

This expansive use of the term incrementality creates ambiguity. When a single metric, such as iROAS, is applied interchangeably across campaigns, baselines, and audience strategies, it can obscure what is actually being measured and make results difficult to interpret or compare. Understanding how incrementality is defined and applied in each context is therefore a critical prerequisite to evaluating measurement rigor and making informed optimization decisions.

Why Does Incrementality Matter?

The advertising measurement landscape is broad, and different metrics answer different questions. However, only causal, incrementality-based measurement allows us to isolate what outcomes occurred because of marketing investment. When advertisers understand incrementality, they can allocate budgets more effectively and maximize true business impact. Relying solely on non-incremental metrics (e.g., ROAS) risks optimizing toward ad exposure and attribution, rather than actual incremental growth.

What Incrementality Question is Being Answered?

Before interpreting iROAS, advertisers should understand exactly what question is being measured. Even when similar methodologies are applied, results can differ materially depending on the unit of analysis, the audience in scope, or the outcome being tracked. In practice, five choices shape what an incrementality result actually represents:

Unit of analysis: Is the analysis conducted at the user, household, store, market, or campaign level?

Exposure being measured: Does the result reflect a specific campaign, an always-on tactic, or a broader audience strategy?

Audience included: Who is in scope for measurement, and was the audience filtered or refined prior to analysis?

Outcome window: What sales or business outcomes are included, and over what time period?

Spend and revenue definition: How are incremental revenue and the corresponding media spend defined within the iROAS calculation?

Two results can both be labeled iROAS and still answer meaningfully different business questions. Clear interpretation requires understanding what is being estimated before evaluating how it was estimated.

How is Incrementality Measured?

There are three common approaches to measuring incrementality present in retail media.

Randomized Controlled Trials (RCTs): Audiences are randomly split into exposed and unexposed groups. This is the most rigorous approach, but it is operationally complex and not always feasible, especially in omni-channel environments.

Matching Methods: An observational approach that estimates impact by building a control group from people or places that didn’t receive ads but look similar to those that did (based on observable factors like past purchases, behavior, or demographics). The assumption is that these “look-alikes” represent what would have happened without advertising. This approach is widely used by RMNs.

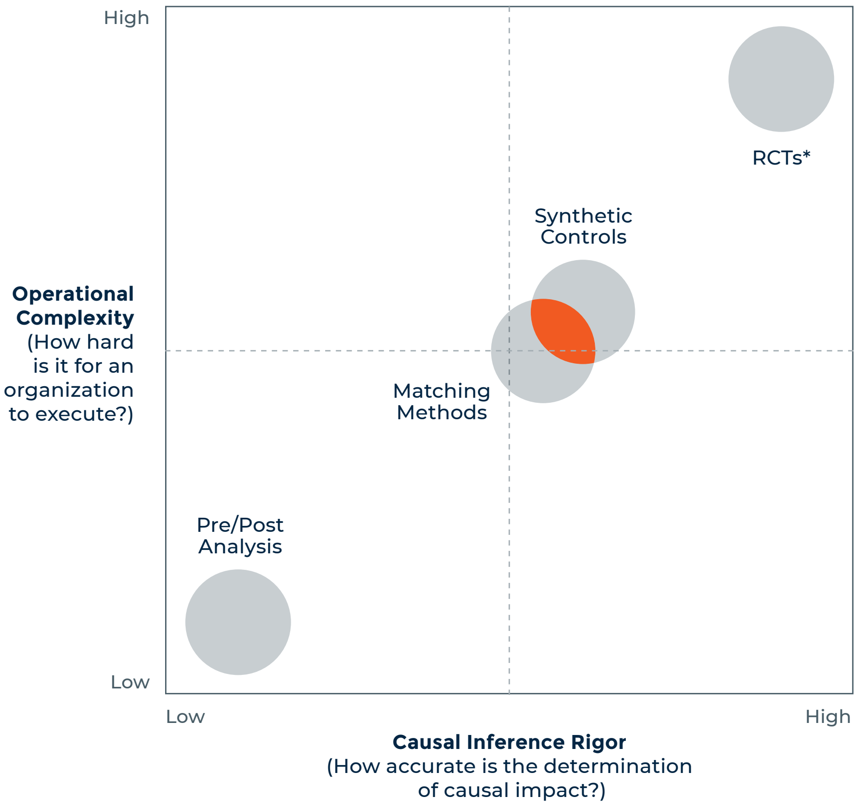
Synthetic Controls: An observational approach that estimates impact by constructing a control group from a weighted composite of markets or users that did not receive ads. Unlike matching, which selects individually similar units based on characteristics, this method finds weights so the composite's pre-campaign trajectory closely mirrors the treated group over time, then projects it forward as the counterfactual during the campaign. This approach is less commonly used by RMNs.

Pre / Post Analysis is another form of measurement that is a basic comparison of results before and after an action that is frequently used in partnership with, or (incorrectly) as a replacement for, incrementality. Pre/post analysis can provide helpful context but should not be used to determine performance or impact.

Each method has a different level of statistical rigor which is connected with the level of operational complexity required to execute. As a result, while some methods enable higher levels of confidence in causal inference outcomes, they may be too operationally complex for a RMN to leverage regularly, or at all ***(Exhibit 1).***

EXHIBIT 1

Operational Complexity vs. Causal Inference Rigor



** Applies only to RCT with randomized test / control groups*

Comparing Matching Methods & Synthetic Controls

The rigor and operational complexity of Matching and Synthetic Controls can vary based on if the test / control units are users or markets as well as data availability.

Synthetic controls require longer historical datasets (~2 years) and can be a better fit for market tests. Matching typically requires less historical data, but that data must enable quality matches which is a better fit for user tests.

Therefore, the rigor and operational complexity of a method will vary based on the retailer, its industry, and available data.

Detailed Overview of Incrementality Measurement Approaches

Each incrementality approach has unique strengths, weaknesses, and complexity in implementation. Understanding these approaches is important for both advertisers and RMNs to ensure proper usage and interpretation of incrementality outputs. Below we outline each incrementality approach in greater detail.

INCREMENTALITY APPROACH 1

Randomized Controlled Trials

Definition

A method of experimentation used to evaluate the impact of a specific ad or campaign by randomly assigning subjects into either a test group (which receives the ads) or a control group (which does not). This randomization ensures that both groups are statistically equivalent at the outset, allowing any differences in outcomes[1] to be causally attributed to the ads themselves. Note, a RCT could explore a variety of situations including incrementality of an individual ad or campaign, head-to-head comparisons of ads, or an increase, decrease, or elimination of ad spend for a specific group. We outline RCTs visually in ***Exhibit 2*** below.

How it Works

In retail media, RCTs are typically executed through:

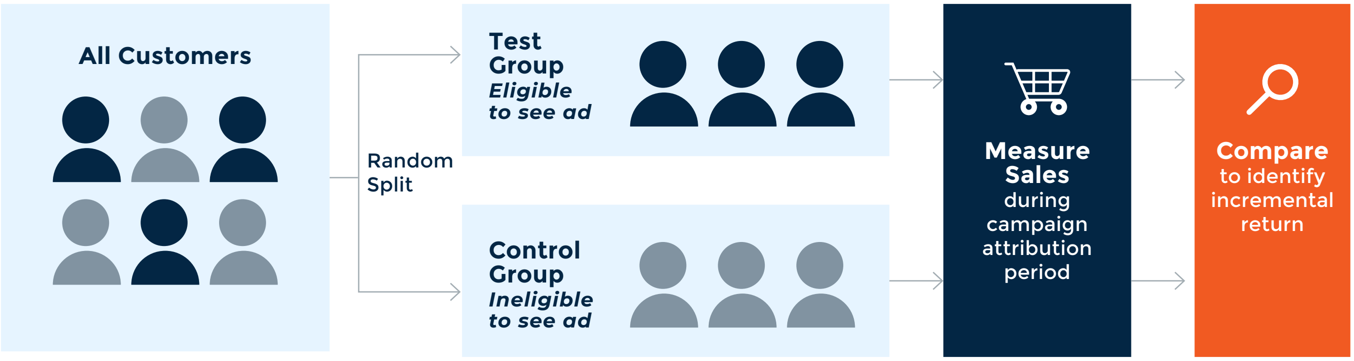
Onsite testing where a broad audience is randomly split on the retailers website into a test group receiving ad exposure and a control group which does not.

Offsite media testing where a broad audience is randomly split in an offsite media platform (e.g., Google DV360) into a test group receiving ad exposure and a control group which does not.

Test and control group sizes are determined through upfront power analysis to ensure statistical validity while balancing media cost and operational constraints.

EXHIBIT 2

RCT Definition


All Customers
Random Split
Test Group
Eligible to see ad
Control Group
Ineligible to see ad
Measure Sales during campaign attribution period
Compare to identify incremental return

[1] If the retailer leverages dynamic pricing, RCTs should be structured to ensure consistent pricing across customers / markets to eliminate contamination.

INCREMENTALITY APPROACH 2

Matching Methods

Definition

A statistical technique used to estimate the causal effect of an ad or campaign by pairing customers or markets in a test group with similar customers or markets in a control group based on observable characteristics (e.g., purchase behavior, demographics). The goal is to approximate the comparison an RCT would provide when random assignment is not feasible and measurement happens after ad delivery is complete.

How it Works

In retail media, the most common forms of matching are:

Cluster Matching: Test and control groups are created by first clustering units based on key covariates (e.g., purchase history, demographics). Then, test units are matched to control units within each cluster to ensure the groups are similar. Key choices include which covariates to use, the clustering method, the number of clusters, and the matching rules used inside each cluster.

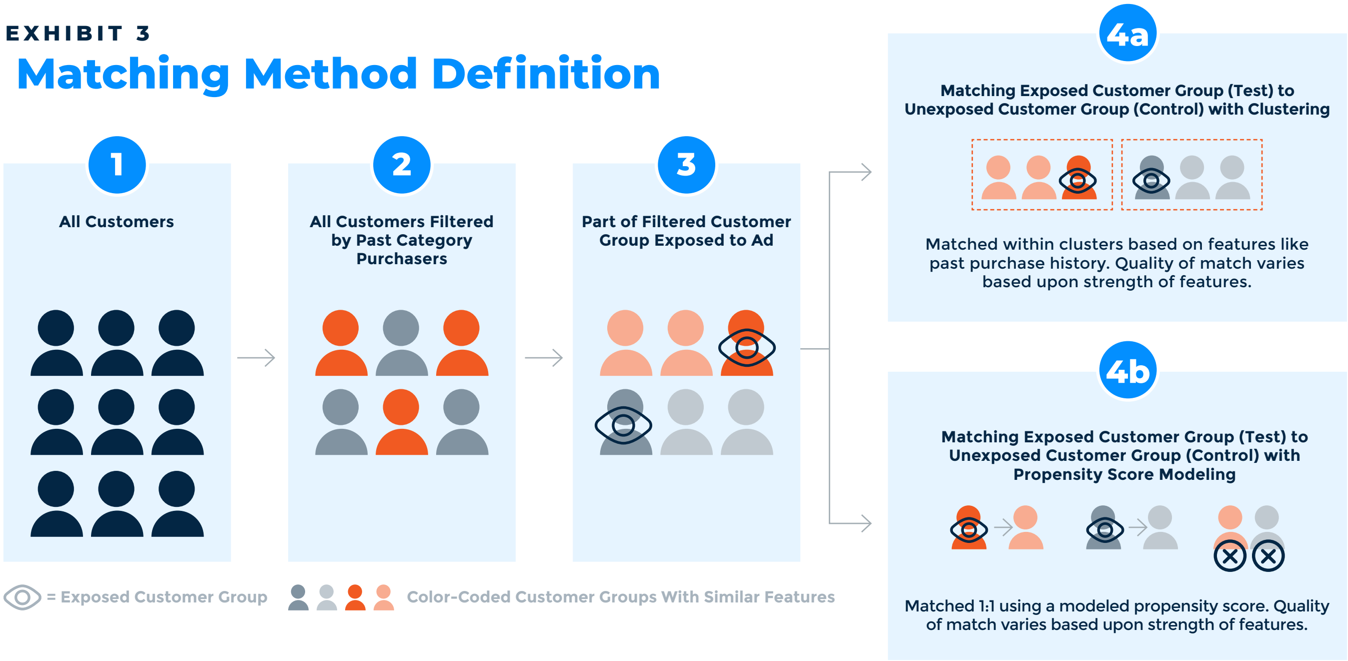
Propensity Score Matching: Test and control groups are matched based on having similar propensity-to-purchase scores. The method can vary based on which covariates are used to calculate the score (e.g., purchase history, demographics) and the matching approach (e.g., K-nearest neighbor).

Before matching, filtering can be done to refine the audience based on different criteria to more closely approximate the likely behavioral characteristics of the test group. For example, an audience might be filtered first to past brand purchasers vs. all customers, and then refined further during matching.

What matching approach is optimal should be driven by data availability and analytical sophistication of the RMN as well as what is important to the advertiser. When creating matched groups the goal is to eliminate both contamination and confounding issues (Examples highlighted in ***Appendix B***).

EXHIBIT 3

Matching Method Definition


1
All Customers
2
All Customers Filtered by Past Category Purchasers
3
Part of Filtered Customer Group Exposed to Ad
4a
Matching Exposed Customer Group (Test) to Unexposed Customer Group (Control) with Clustering
4b
Matching Exposed Customer Group (Test) to Unexposed Customer Group (Control) with Propensity Score Modeling
= Exposed Customer Group
Color-Coded Customer Groups With Similar Features

Matched within clusters based on features like past purchase history. Quality of match varies based upon strength of features.

Matched 1:1 using a modeled propensity score. Quality of match varies based upon strength of features.

INCREMENTALITY APPROACH 3

Synthetic Controls

Definition

An observational approach that estimates impact by constructing a control group from a weighted composite of markets or users that did not receive ads. Unlike matching, which selects individually similar units based on characteristics, this method finds weights so the composite's pre-campaign trajectory closely mirrors the treated group over time, then projects it forward as the counterfactual during the campaign.

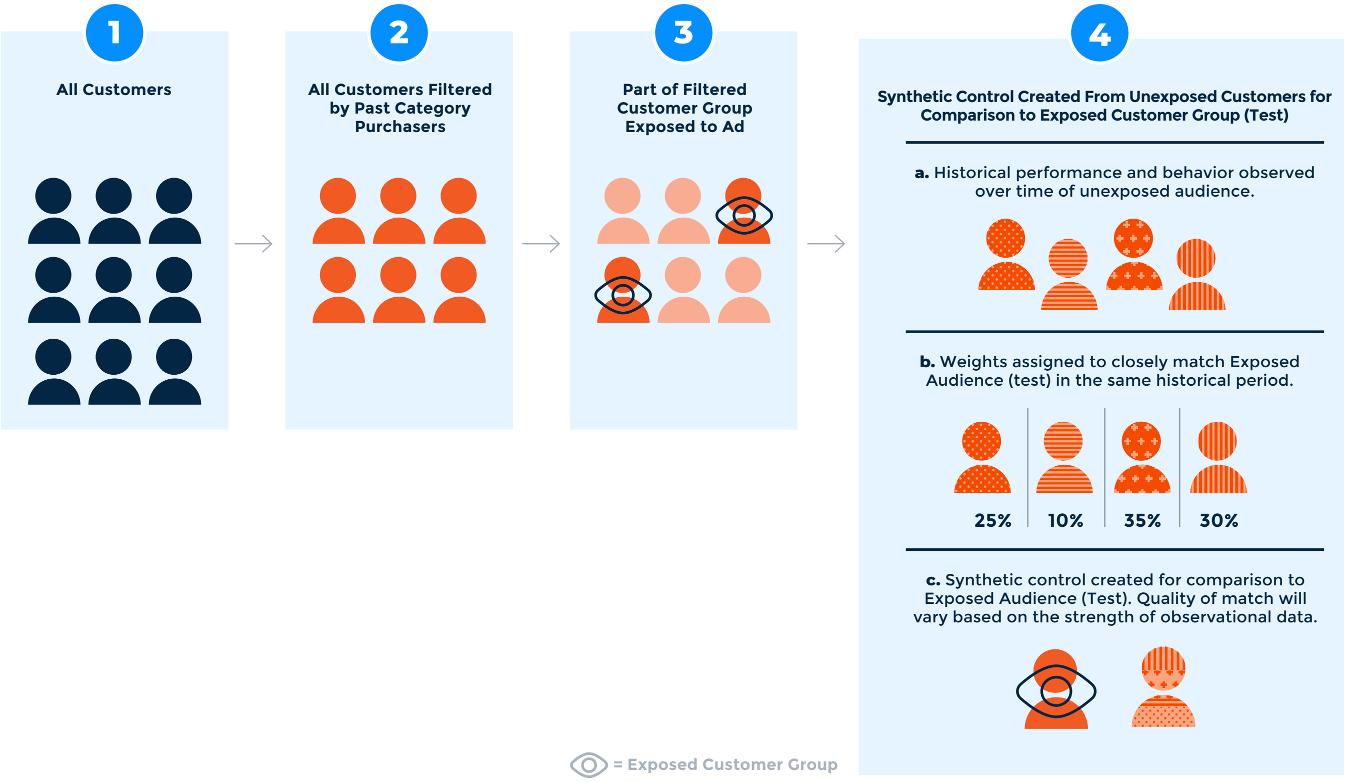
How it Works

The method uses pre-period outcome data from a pool of untreated units and assigns weights to create a “synthetic” control that best matches the treated unit before the campaign. If the synthetic control tracks the treated unit closely in the pre-period, it provides a credible estimate of what would have happened without advertising. Because this approach typically relies on a long history of outcomes (often ~1–2 years), it is generally best suited for geo/market-level tests rather than individual-level user matching.

Like matching, filtering may also be done before creating the synthetic control to more closely approximate the likely profile of the test group (e.g., past brand purchasers).

EXHIBIT 4

Synthetic Control Definition



1 All Customers

2 All Customers Filtered by Past Category Purchasers

3 Part of Filtered Customer Group Exposed to Ad

4 Synthetic Control Created From Unexposed Customers for Comparison to Exposed Customer Group (Test)

a. Historical performance and behavior observed over time of unexposed audience.

b. Weights assigned to closely match Exposed Audience (test) in the same historical period.

c. Synthetic control created for comparison to Exposed Audience (Test). Quality of match will vary based on the strength of observational data.

= Exposed Customer Group

Approach to Calculating Incremental Revenue between Control and Test Groups

Once a control group is constructed, using either matching or synthetic control approaches, sales are isolated in one of two ways enabling comparison between the test and control groups to determine incremental revenue.

Observed Sales Performance

The actual sales of the matched or synthetic control group and test group are leveraged for comparison during the same period as the test group's ad exposure. A t-test is then used to determine statistical significance.

Bayesian Structural Time Series (BSTS)

Is a more complex approach where a time series causal inference model (e.g., Google's Causal Impact) estimates what would have happened to the test group in the absence of ad exposure by learning historical relationships between the test group, control group, and time patterns. The modeled test group sales is then compared to the control group to determine incremental revenue. This approach explicitly accounts for trends and seasonality. Note, BSTS provides credible intervals to understand distribution of the lift vs. using a t-test with observed sales performance.

There are strengths and weaknesses to each of these approaches which are outlined in ***Appendix C***. The optimal approach should be one that best matches the context of the analysis but also that is used consistently across similar analyses.

Incrementality Measurement in Retail Media

In the retail media landscape, the approach to incrementality measurement is driven by the sales channel focus of the retailer (e-commerce, omni-channel) and the overall sophistication of its measurement capabilities (data availability, advanced tech stack, modeling expertise).

E-commerce

Retailers with a concentrated business online have a greater ability to offer RCTs to measure incrementality. The most sophisticated retailers are able to offer self-service capabilities for creating randomized test and control groups and launching RCT tests in-market.

Omni-Channel

Retailers with a mix of online and store sales have less ability to offer RCTs given the multiple ways customers can be exposed to ads and then shop. It is much more difficult to maintain clean test and control groups, and as a result, they are more likely to rely on post-campaign observational approaches to offer incrementality measurement. The production and analysis of the models are typically done in-house by the RMN and aggregated results are provided to the advertiser. The sophistication of matching approach and test / control group comparison typically aligns to the size of the retailer.

If a RMN isn't using an RCT to measure incrementality, then they are likely using a matched control and less frequently a synthetic control group. As highlighted, matching and synthetic control approaches can vary substantially. Simple choices in methodology can generate meaningfully different outputs and potentially different choices by advertisers.

To demonstrate this challenge, we reviewed 42 campaigns within Albertsons Media Collective (The Collective) to demonstrate how different choices within its matching approach could drive differences in iROAS.

Our analysis, outlined in Part 2 of this whitepaper, found that:

6.5X

Variation in iROAS outputs

(median 2.5x)

83%

of campaigns could flip from positive to negative based on methodological choices

Part 2

Analysis of Matching Approaches in 42 Digital Ad Campaigns

Bringing Incrementality Measurement to Life | Analysis of Matching Approaches

As an omni-channel retailer, The Collective cannot always rely on RCTs to measure campaign incrementality at scale. Instead, The Collective often uses matching methods to construct control groups for comparison against ad-exposed customers.

To demonstrate how different matching approaches could impact iROAS outputs, we varied four matching methodology options across 42 different campaigns (Exhibit 5). Campaigns were representative across categories, time periods, campaign objectives, and brand size. All campaigns evaluated were onsite display ads across Albertsons website properties (e.g., Albertsons.com, Vons.com, Safeway.com).

EXHIBIT 5

Matching Methodology Varied in Analysis

	Methodology Choices Across iROAS Analysis	Options
1	**Before matching, should the test and control group be FILTERED in any way?**	• Past brand customers • Past category customers • No filter
2	**What MATCHING APPROACH should be taken?**	• Clustering • Propensity Score Matching (PSM)
3	**Within my chosen method, what FEATURES and algorithms do I use to match?**	**Features for Clustering** • Division • Category sales • Brand sales • Digital engagement • Site sessions **Clustering Algorithm** • K-means **# Neighbors for PSM using K-nearest neighbors** • 1:1 • 1:Many
4	**How do I isolate sales and CALCULATE INCREMENTAL REVENUE between my control and test group?**	• Observed Sales Performance with t-test • Bayesian Structural Time Series (BSTS)

The different methodology choices outlined result in a range of 54 different iROAS values without changing anything about the structure of the campaign itself.

Analysis Findings

Our analysis varied matching methodology choices at each step of calculating iROAS across 42 campaigns. As noted above, these steps included:

1 Test and control group filtering

2 Matching approach selection

3 Feature selection for matching using brand sales

4 Approach to calculating incremental revenue

We then compared the range of iROAS outputs and the quality of matches to understand how methodological choices can impact perceived performance of a campaign.

Test and Control Group Filtering

Exhibit 3: Step 1

Before matching, we explored the impact of filtering both the test and control group. Initial filtering is typically done to make test and control groups more similar to ensure measurement is performed on like-for-like groups.

Our analysis filtered the initial test and control group to customers with previous category sales and brand sales. As demonstrated in Exhibit 6, this reduced the average sample size of our test and control group somewhat for category sales (-17%) and substantially for brand sales (-83%). When applying these filters, we found the average quality of match (as defined as Standard Mean Difference[2]) between the test and control group worsened for a category sales filter (7%) and meaningfully (35%) for a brand sales filter.

When applying a brand sales filter, we also found that iROAS worsened substantially ($2.27 to $0.22) while a category sales filter remained more stable ($2.27 to $2.20). As we narrowed our audience with a filter, we also narrowed incremental sales (numerator) while keeping advertising spend unchanged (denominator). Given their hierarchy, a category sales filter will always have more incremental sales than a brand sales filter.

Key Learning

Filtering can allow advertisers to better understand the impact of their ads on a specific audience (e.g., category buyers who aren't brand buyers). However, filtering can dramatically impact sample size and reduce iROAS independent of ad effectiveness. The reduction of sample size may also impact match quality, either positively or negatively depending on the specific nuances of the audience. The right filtering approach should balance maintaining sample size while striving to create a control group as comparable to the test group as possible.

[2] Standard Mean Difference (SMD): The difference in group means divided by the pooled standard deviation (SMD = $(\text{mean}_1 - \text{mean}_2)/s$. For our purposes, SMD summarizes how similar two groups are, and thus the quality of their match where lower (closer to 0) is better.

EXHIBIT 6

Control Group Filtering Impact on Sample Size

	Filter to Customers with Brand Sales in Prior Year	Filter to Customers with Category Sales in Prior Year	No Filter
Avg Number of Customers in Test Group *(Indexed to No Filter)*	17	84	100
Avg Number of Customers in Control Group *(Indexed to No Filter)*	16	81	100
Avg % of Test Group with $0 Sales of Brand in Prior Year	0%	82%	85%
% of Test Converters that are New-to-Brand[2]	0%	38%	41%
Avg Standard Mean Difference of Relevant Covariates	0.101	0.080	0.075
Average iROAS in Aggregate[3]	$0.22	$2.20	$2.27

[2] New-to-Brand defined as those who saw the ad and purchased during the campaign but had not purchased the brand in the 52W prior to the campaign

[3] Average iROAS in Aggregate = Average of iROAS across all campaigns using two incremental revenue calculation approaches (1) Observed Sales Performance and (2) Bayesian Structural Time Series (BSTS).

Matching Approach Selection

Exhibit 3: Step 4a and 4b

We evaluated the impact of two matching approaches, Propensity Score Matching (PSM) and Clustering. Within each approach, we assessed performance across a range of feature and filter combinations (see Feature Selection for Matching). For the initial comparison of PSM versus Clustering, we aggregated metrics across all feature and filter combinations.

Our analysis found that PSM delivered, on average, an approximately 12x lower SMD across relevant covariates compared to clustering (***Exhibit 7***), indicating a substantially higher-quality match and that differences in observed covariates are not driving the result. This confidence is demonstrated through the distribution of SMD results which is more narrow for PSM vs. Clustering (***Exhibit 8***).

Additionally, PSM resulted in fewer unique matched control customers than Clustering due to the stricter similarity requirements inherent to the method. This effect was most pronounced when using PSM 1:1 compared to PSM 1:many. PSM and Clustering produced different aggregate iROAS estimates (Clustering: $1.80 vs. PSM 1:many: $0.73).

Key Learning

Each matching approach has tradeoffs. While PSM delivers stronger covariate balance and greater confidence in causal estimates, it typically yields fewer unique matched control customers and may be associated with lower iROAS estimates. Additionally, depending on the RMN, PSM can be more operationally complex to implement.

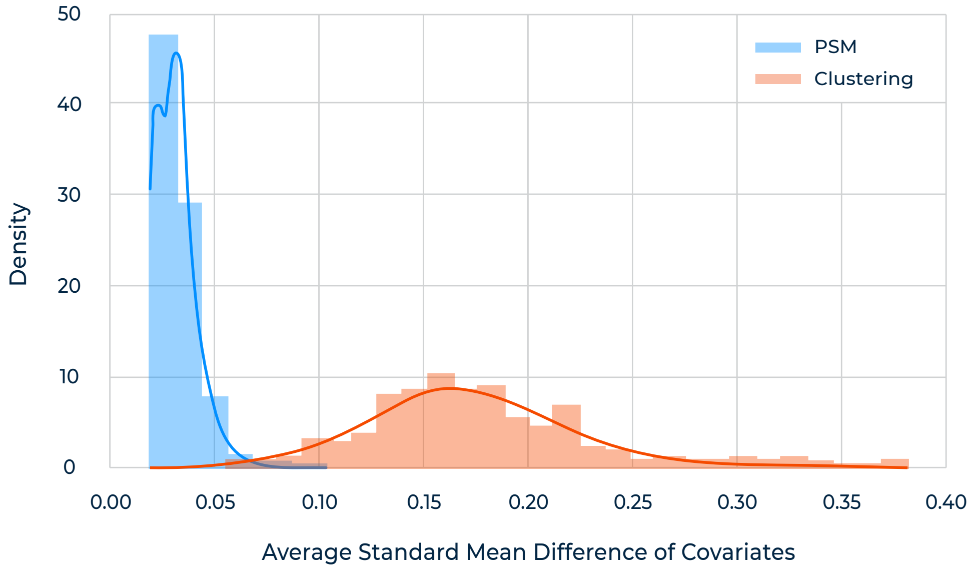
EXHIBIT 7

Quality of Match Across Features in PSM and Clustering

Across all filter/feature variations	Clustering	Propensity Score Matching (1:Many)	Propensity Score Matching (1:1)
Avg Number of Customers in Matched Control Group *(Indexed to itself)*	100	100	100
Avg Number of Customers in Unique Matched Control Group *(Indexed to Avg Customers in Matched Control Group)*	100	22	29
Average Standard Mean Difference of Covariates	0.158	0.013	0.012
Average iROAS in Aggregate	$1.80	$0.73	$0.84

EXHIBIT 8

Distribution of SMD across PSM and Clustering



Feature Selection For Matching Using Brand Sales

Exhibit 3: Step 4a and 4b

Next we explored how different features for matching impacted the quality of matches and performance. Within each approach, we examined the impact of matching with and without past brand sales. The other features included for matching were digital engagement (e.g., online ordering, digital coupon usage), number of digital sessions (i.e., site / app visits), and division (e.g., Denver, Northern California).

When brand sales are not used for PSM (1:many), the average iROAS changes from $1.23 to -$0.14. When Clustering, omitting brand sales results in an average iROAS that holds steady at $1.81 (***Exhibit 9***). This is driven by the underlying approach for Clustering where a feature change that is related to the outcome (i.e., brand sales) may shift clusters slightly, but test and control groups remain fairly similar. However, with PSM when a feature is removed the underlying modeled propensity score shifts creating less similar test and control groups.

Key Learning

A best practice in matching feature selection is to identify features with the best correlation to, or likelihood of predicting, the intended outcome (sales). For The Collective, this was past brand sales and therefore, use in feature selection helped ensure the similarity of test and control groups.

We anticipate that many other RMNs focused on frequency based sales would similarly find past brand sales to be the optimal feature for matching. However, some RMNs may find that other features are better predictors given the dynamics of their industry and customer (e.g., loyalty membership).

A final learning was the sensitivity of PSM to changes in feature selection. RMNs who leverage PSM should ensure that feature selection for matching remains consistent across brands and campaigns to minimize volatility in iROAS outputs.

EXHIBIT 9

iROAS Across Features Between Clustering and PSM

	Propensity Score Matching		Clustering	
	With Historical Brand Sales as a Feature	WITHOUT Historical Brand Sales as a Feature	With Historical Brand Sales as a Feature	WITHOUT Historical Brand Sales as a Feature
Average iROAS in Aggregate	**$1.23**	**$-0.14**	**$1.81**	**$1.81**

Approach to Calculating Incremental Revenue

Exhibit 3: Step 4a and 4b

Lastly, we compared incremental revenue calculated using an Observed Sales Performance approach versus a Bayesian Structural Time Series (BSTS) model. We found that BSTS can produce materially different incrementality results compared to Observed Sales Performance. This divergence is expected, as BSTS explicitly accounts for time-series dynamics such as trend and seasonality, while observational iROAS is typically point-in-time and only reflects these patterns when they are manually controlled.

Across our analysis, the average difference between the two iROAS estimates was ~90% ($1.56 Observed Sales Performance iROAS vs. $0.97 BSTS iROAS). In ***Exhibit 10***, we also found meaningful differences between the two iROAS estimates across different matching approaches (PSM / Clustering) and feature selection (with and without brand sales as a feature).

Key Learning

Observed Sales Performance and BSTS can tell very different stories in iROAS outputs. In general, BSTS results were lower and had less variation than Observed Sales Performance. BSTS's approach of using past seasonality and trends in determining incremental revenue likely supports a more rigorous iROAS across methodologies. However, BSTS usage does require a more sophisticated data science team and toolkit which may not be feasible for all RMNs.

EXHIBIT 10

Average iROAS Across Variations: Observed Sales Performance vs. BSTS

	Avg Observed Sales Performance iROAS	Avg BSTS iROAS	Avg % Diff
Across All Variations of Clustering & PSM (1:Many)	**$1.56**	**$0.97**	**90%**

Exhibit Note: Only considering PSM 1:Many so even split between two methods – 252 total – 126 from each method
Exhibit Note: % difference calculated at the campaign/variation level: (iROAS - BSTS iROAS)/BSTS iROAS

EXHIBIT 11

Average iROAS Across Variation Using Observed Sales Performance vs. BSTS

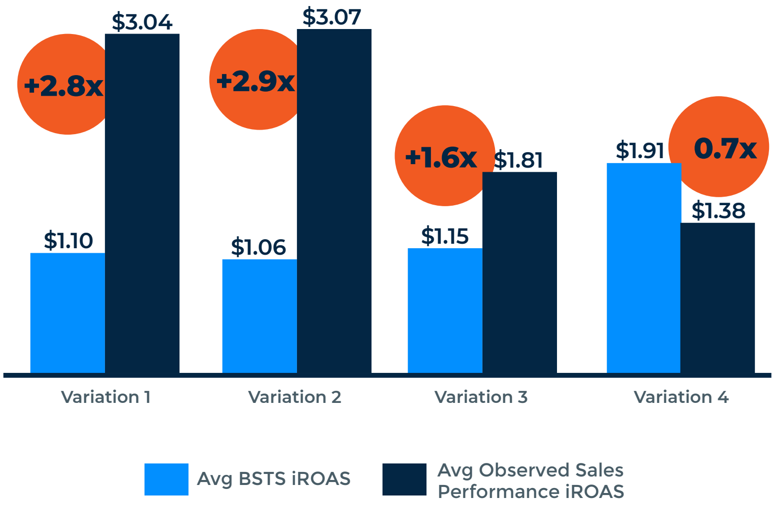


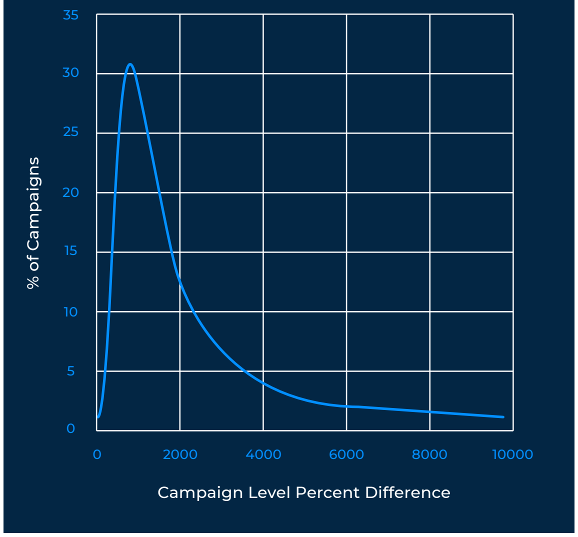
Exhibit Note: All variations include no filtering. Variations include PSM (1:many) or Clustering with and without Brand Sales as a feature.

Analysis Summary

By varying matching methodologies, our analysis produced 54 distinct iROAS outcomes across 42 campaigns. Within individual campaigns, the gap between the highest and lowest iROAS averaged 6.5x (median 2.5x), and 83% of campaigns flipped from positive to negative depending on methodology (***Exhibit 12***). These swings demonstrate that observed performance is often as influenced by methodological decisions as by underlying customer behavior, making it critical to understand how matching choices shape results.

EXHIBIT 12

Distribution of Campaign Level % Difference Between Highest and Lowest iROAS Generated



From this analysis, we identified five key takeaways to understanding the impact of matching methodological choices on outcomes.

1

iROAS is highly sensitive to methodology. Across filtering, matching approach, feature selection, and incremental revenue calculation, the same campaign can produce a wide range of iROAS outcomes driven by measurement choices and not consumer response.

2

Filtering improves audience specificity but can weaken the measurement. Filtering to past category/brand purchasers reduces the available control pool, which lowers sample size, impacts match quality (positively and negatively), and often depresses iROAS, independent of true ad effectiveness.

3

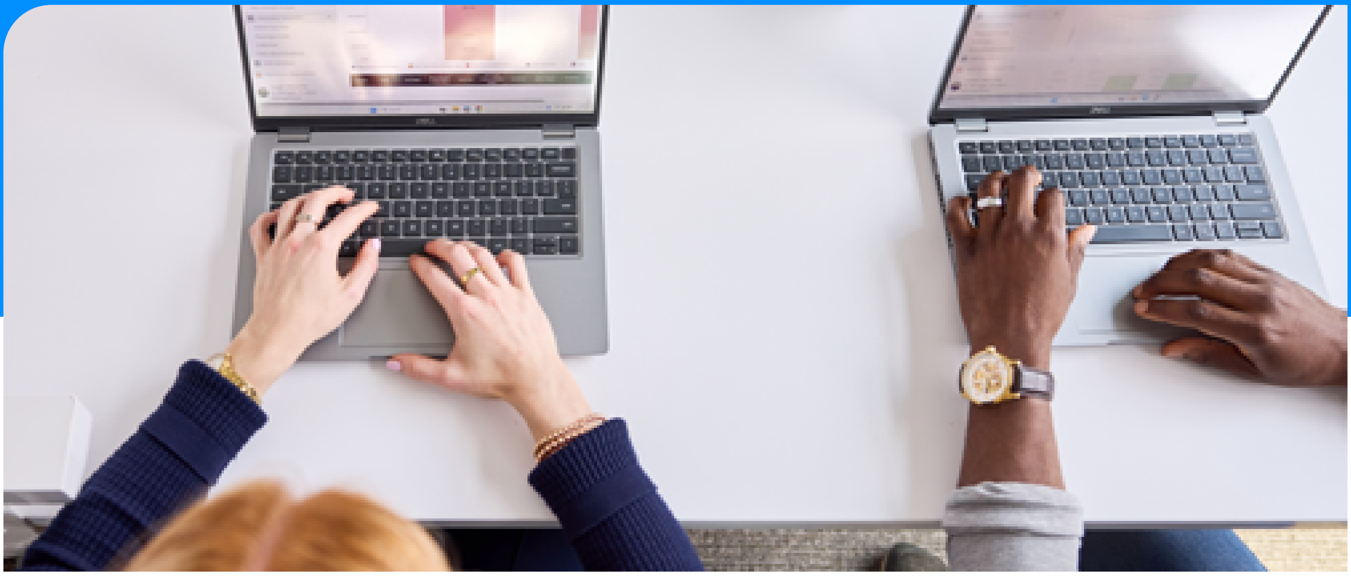
PSM produces stronger matches and more consistent results than Clustering. PSM generally delivers much better covariate balance and lower output variance. Clustering tends to produce weaker matches and higher iROAS, suggesting a greater risk of overstating incremental revenue.

4

Feature selection is a key lever, especially for PSM. In the case of The Collective, including historical brand sales materially improves match quality and stabilizes outputs. Removing brand sales from the feature set can dramatically shift PSM estimates, while clustering is less sensitive.

5

BSTS and Observed Sales Performance often diverge meaningfully. BSTS typically yields lower variability than Observed Sales Performance because it does a better job of controlling for confounding variables like sales trends and seasonality.



What This Means for Advertisers

Given the wide range of incrementality methodologies, advertisers often struggle to interpret, trust, and compare incrementality results, both within a single RMN and across partners. Ultimately, there is no single “right” way to measure incrementality. Instead, our evaluation of methods and analysis highlights three principles advertisers should prioritize:

Transparency

Insist that RMNs clearly document every step of their incrementality approach—who is included, how test/control groups are constructed, what features are used, and how incremental revenue is calculated. Without transparency, results cannot be evaluated or trusted.

Consistency

Maintain consistent incrementality approaches over time within each RMN, and align approaches across RMNs where possible to improve interpretability. **Note:** even when two RMNs use the same incrementality method, results should not be directly compared or summed, since attribution logic, extrapolation, and measurement scope often differ.

Understanding

Ensure day-to-day users understand the methodology well enough to interpret results correctly. This is especially true of assumptions, limitations, and situations where the method may over or under-state impact. Incrementality only drives better decisions when stakeholders understand what the metric truly represents.

Bringing it All Together

Retail media has reached a critical inflection point. Incrementality, and iROAS specifically, has become central to how advertisers evaluate performance, allocate budgets, and justify investment. Yet, as this research demonstrates, materially different incrementality outcomes can be produced for the same campaign based solely on methodological choices that are often opaque to advertisers.

This lack of transparency creates risk for all stakeholders. Advertisers struggle to interpret and compare results, RMNs face growing skepticism around reported performance, and the industry risks eroding trust in one of its most important performance metrics.

To help address these challenges, this research offers practical guidance for both sides of the ecosystem. Advertisers will find a diagnostic tool in ***Appendix D*** — a set of structured questions designed to support iROAS understanding and help evaluate the methodological rigor behind any incrementality measurement they receive. For RMNs, ***Appendix E*** provides actionable recommendations to consider as they build and mature their measurement capabilities.

Together, these resources are intended to move the industry toward a more transparent, comparable, and trustworthy incrementality measurement standard — one where methodology is not a black box, but a shared foundation for confident decision-making.

Industry Call to Action: Standards for Transparency

Retail media and advertiser industry groups are uniquely positioned to address the challenges we have identified in our research. We believe the next phase of retail media maturity requires industry-wide transparency standards for incrementality measurement, not to mandate a single methodology, but to ensure results are interpretable, comparable, and trusted.

EXHIBIT 13

Proposed Transparency Standards for Incrementality Measurement in Retail Media

Area	Recommended Transparency Standard	Why This Matters
Incrementality Method Classification	• Identify the primary incrementality approach used (e.g., RCT, matching-based observational method, synthetic control, modeled approach) • Note when multiple or hybrid methods are used	Incrementality approaches vary in causal rigor. Clear classification enables advertisers to properly interpret results and understand differences across RMNs and campaigns.
Test and Control Group Construction	• Specify the unit of analysis (e.g., user, household, store, geo) • Disclose any audience filtering applied prior to measurement • Describe control group construction and known sources of contamination or bias	Filtering and control group design can materially shift iROAS outcomes by affecting sample size, match quality, and incremental revenue—independent of true ad effectiveness.
Matching or Modeling Methodology	• Describe the matching or modeling approach used • List primary features or inputs • Explain how match quality or model fit is evaluated • Indicate whether methods are applied consistently across campaigns	Small methodological choices can drive large swings in iROAS. Transparency allows advertisers to assess stability, bias risk, and consistency of reported performance.
Incremental Revenue Calculation Approach	• Clarify whether observed sales or modeled approaches (e.g., BSTS) are used • Describe how trend and seasonality are handled • Disclose measures of statistical confidence or uncertainty	Observed and modeled approaches can diverge materially. Clear disclosure is critical to avoid misinterpretation and suboptimal optimization decisions.
Appropriate Use Cases and Limitations	• Provide guidance on when results are most reliable (e.g., spend levels, campaign types) • Call out known limitations • Offer guardrails for interpretation and use	Clear guidance improves decision quality, prevents misuse, and protects advertiser trust.

Appendix

Strengths and Weaknesses Across Incrementality Measurement Approaches

Incrementality Measurement Approach	Strengths	Weakness
Randomized Controlled Trials (RCTs)	• Audience-based tests • High rigor for audience-based tests • Strong rigor for geo-based tests but higher risk of contamination vs. audience-based tests • Best in class approach for determining causal inference • Easy to interpret	• Time-intensive and operationally complex requiring pre-design to maintain clean test and control groups • Potential audience / revenue loss due to the holdout of media on control group • Prone to data contamination in control groups • Large sample size required to reach statistical significance which may not be practical for small brands / RMNs / campaigns
Matching (Clustering and PSM)	• Highly scalable • Allows for mid-campaign optimizations • Can be done retrospectively (after a campaign) • Enables more granular insights	• Susceptibility to hidden model bias (e.g. omitted variable bias, selection bias) • Quality of matches can be poor if test group is very different from the control group
Synthetic Controls	• Highly scalable • Allows for mid-campaign optimizations • Can be done retrospectively (after a campaign) • More transparent modeled approach compared to matching	• Requires robust historical and granular customer data if done for user testing • Model complexity

Appendix: B

Matching Contamination and Confounding Issues Examples

	Definition	Example	Why This Happens
Contamination	When customers are included in a control group when they were unintentionally exposed to the test ad	Customers in a control group that were exposed to ad on a different device	• Customer data is not connected across devices • Test audiences are exposed to a brand's ad in a different environment
Confounding Issues **(Selection Bias)**	When factors unintentionally bias comparison between a test and control group making it unclear if an observed effect is caused by the ad	Customers included in the control group who have unique behavior based on demographics or seasonality	• Features for filtering and matching do not align to ad's objective (e.g., control group is filtered to older age demo for diaper ad)

Strengths and Weaknesses of Approach to Calculating Incremental Revenue between Control and Test Groups

	Observed Sales Performance	Bayesian Structural Time Series (BSTS)
Strengths	• Simple, fast, transparent • Easy to communicate and reproduce • Fewer modeling decisions	• Models trend and seasonality explicitly • Can use control series to reduce bias • Produces credible intervals and probability of lift
Weaknesses	• Does not account seasonality / trends • Assumes independent observations • Produces a single average effect vs. time-varying impact	• More complex to specify, QA, and explain • Needs enough pre-period data to learn patterns • Can feel like a "black box" to stakeholders
When to Use it	Simplicity and ease of communication matters and seasonality is less of a concern	Seasonality is important to the analysis and consumers of outputs have a strong analytical foundation

Advertiser Question Guide

A structured guide for evaluating and interpreting incrementality results from Retail Media Networks

Category	Key Questions
Scope & Definition	• What exactly is this result measuring? Is this iROAS for a specific campaign, an always-on tactic, or a defined audience segment? • What unit is being measured — user, household, store, market, or something else? What sales window and outcome definition are included? • Who is included in the analysis? Was the audience filtered before measurement? Who was excluded, and why? • How does that filtering change the business question this iROAS is answering?
Method Understanding	• What is the RMN's approach to determining incrementality and which methodology is being used (e.g., RCTs, Matching, Synthetic Controls)? • Who oversees incrementality analytics at the RMN and have methodological approaches been documented to share with advertisers? • What level of spend or campaign type does the RMN recommend iROAS measures are most accurate? • At a high level, how was the control group built — was the analysis based on an RCT, matching, synthetic control, or another modeled approach?
RCTs	• How does the RMN randomize groups and ensure they're balanced? • How is the RMN detecting and correcting for contamination or bias in sample selection?
Matching Methods	• What features are used to match on? • What matching approach is used (e.g., Clustering, PSM)? • How is the quality of the match evaluated? • How is incremental revenue calculated between test and control groups (e.g., BSTS or observed performance)? • How were seasonality, trend, and baseline demand accounted for in the revenue calculation? • If observed performance is used, what is your significance threshold for providing results? What measures of uncertainty or confidence accompany the result? • How is the RMN detecting and correcting for contamination or bias in sample selection?
Synthetic Controls	• What features are the synthetic control built on (e.g., user, geos)? • How is incremental revenue calculated between test and control groups (e.g., BSTS or observed performance)? • How were seasonality, trend, and baseline demand accounted for in the model? • If observed performance is used, what is your significance threshold for providing results? What measures of uncertainty or confidence accompany the result?

Continued on page 31

Advertiser Question Guide

Continued from page 30

Category	Key Questions
Comparability & Limitations	• Can this result be compared to prior campaigns on this RMN? Has the methodology changed over time in ways that affect comparability? • Is this number being compared against results from another RMN that may use a different scope, baseline, or methodology? • For what campaign types, spend levels, or retail environments is this methodology most reliable? • What are the known limitations or situations where the result may over- or under-state incremental impact?
Decision Support	• What decision is this result intended to support — budget allocation, optimization, or partner comparison? • Is this result strong enough to guide that decision, or does it require additional context before acting? • What additional information should be reviewed alongside this iROAS before making a business decision?

What This Means for RMNs

Given our recommendations for advertisers, RMNs should prepare for a greater push for transparency, consistency, and understanding in incrementality measurement. To enable these needs RMNs should prioritize five key areas in their measurement capability roadmaps.

Document Incrementality Approaches for External Consumption	• Clearly documented incrementality approaches and methodology in formats that can be easily shared with advertisers.
Train Sales and Product Teams on Approaches	• Sales and product teams should fluently understand how incrementality methodology works to enable productive conversations with advertisers that drive understanding and trust.
Invest in Data and Analytics Tools To Enable Measurement	• Build robust data architecture that collects customer data (e.g., purchase history, digital engagement, propensity scores, demographics) that enables incrementality measurement • Invest in architecture that enables analytics (e.g., data platform / compute, modeling / transformation, automation etc.)
Invest in Data Science and Analytics Expertise	• As demand for more robust and varied incrementality measurement increases from advertisers, RMNs will need to build in-house teams that have the know-how and capacity to support.
Plan for Flexibility	• Anticipate that sophisticated advertisers will want incrementality measurement done based on specific approaches to enable better cross-RMN comparison. RMNs with robust data and analytics tools and teams should plan to flexibly meet these needs for their high-spend advertisers

Author Bios

Albertons Media Collective

Sophie Armor

is a Data Scientist at Albertsons Media Collective. She brings both financial services and retail media expertise to her work and specializes in causal inference and insight generation, leveraging her expertise to drive impactful data-driven decisions.

Levi Dantzinger

is the leader of Marketing Science Research at Albertsons Media Collective. He brings analytics and data science experience across industry verticals including CPG, Defense, Healthcare, InsureTech, and retail media. At The Collective, he focuses on innovating measurement methodologies and metrics to help advance causal understanding of retail media's impact on buyer behavior to help drive meaningful incremental value for brands.

Northwestern University's Kellogg School of Management

Eric T. Anderson

is the Polk Bros. Chair and Professor of Marketing at Northwestern University's Kellogg School of Management. He has served on the board of directors at Canadian Tire since 2016. His research and advising focus on applied analytics and ML/AI in retail, eCommerce and financial services.

Brett R. Gordon

is the Charles H. Kellstadt Professor of Marketing at Northwestern University's Kellogg School of Management. His research focuses on helping firms optimize their pricing, promotion, and advertising strategies through analytical modeling and field experiments. His current work addresses the challenge of measuring incremental advertising effectiveness while accounting for real-world marketing constraints.

Ovative Group

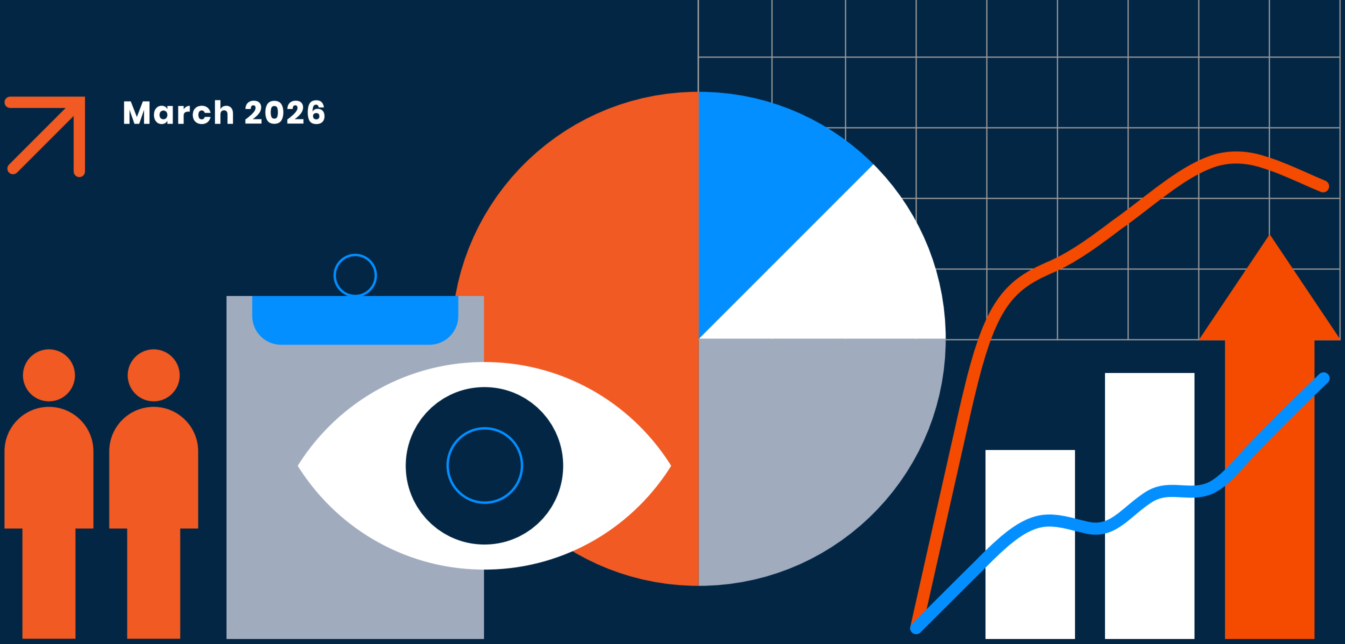
Kate Bante

is Vice President of Measurement at Ovative Group. She leads a team of marketing science and analytics experts focused on helping clients address key business questions through advanced modeling and testing. Additionally, she leads measurement initiatives tackling complex industry challenges, including brand and retail media measurement.

Derek Nelson

is a Senior Director of Retail media at Ovative Group. He leads a retail media Consulting team that guides mature and emerging commerce media networks to grow revenue, create efficiencies, and delight their internal and external partners. With extensive experience in leading retail media networks, Derek specializes in network measurement, reporting, and data monetization.

March 2026



Have a question or want to stay connected?

→ marketing@ovative.com

ovative

Albertsons Media COLLECTIVE

In partnership with professors from Northwestern University Kellogg School of Management.

marketing@ovative.com | mediacollective@albertsons.com

ovative.com
albertsonsmediacollective.com