Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

From Attributed to Caused: Operationalizing Always-on Audience Holdouts for Omnichannel Incrementality in Big-Box Retail

Venkata Suneel Dasika

Independent Researcher, India. Email: gowthampappulaconnects@gmail.com

Abstract
Advertising measurement across networks is difficult to compare because platforms use different exposure definitions, attribution windows, identity rules, and counterfactuals. This paper describes an operational protocol for always-on audience holdouts in a large omnichannel retail setting. Eligible first-party customers are assigned by stable randomization to a persistent portfolio holdout or the addressable media population; paid delivery is suppressed to the holdout as completely as platform capabilities permit, and outcomes are read from retailer-owned omnichannel transaction data, including in-store sales. The primary causal estimand is the intent-to-treat difference in sales between randomized populations, not a customer-level propensity residual or a platform-reported conversion count. Match rate, suppression leakage, and exposure timing are treated as compliance and interpretation diagnostics rather than substitutes for randomization. The global holdout identifies portfolio-level incrementality. It does not, by itself, identify additive channel contributions; channel-specific causal effects require separate randomized channel holdouts under the same outcome and read-window rules. Channels that cannot support person-level suppression are evaluated through geo experiments or remain outside the strict protocol. Because suppression can be imperfect, estimates may be attenuated toward zero under stated assumptions and are reported with confidence intervals and leakage diagnostics. No quantitative outcomes are reported because the paper focuses on the design and governance of the framework rather than an evaluation of a named deployment. The contribution is a methods and governance architecture for making omnichannel incrementality reproducible and retailer-owned.

Keywords: paid media incrementality, incremental return on ad spend, randomized audience holdout, intent-to-treat estimation, portfolio incrementality, first-party identity graph, suppression compliance, exposure leakage, omnichannel sales measurement, multi-touch attribution limitations, geo-lift testing, cross-network measurement governance, measurement reproducibility, privacy-durable measurement.

I. Introduction

1.1 A decision problem, not a reporting problem

A retailer allocating paid media budget across search, social, programmatic, digital audio and video, connected television, and retail media has to answer one question repeatedly: if this spend were removed, what would happen to sales? Every downstream decision (budget shifts, channel commitments, forecasts given to the business) rests on that answer.

The measurement systems available to answer it do not agree, and the disagreement is not noise. Each network answers a slightly different question and reports the answer as though it were the same one. They differ in what counts as an exposure, in the credit window applied, in how identity is resolved, in what is observed versus modeled, and, most consequentially, in how the counterfactual is constructed. Since a lift figure is a difference against a baseline, and the baselines differ, the figures are not on a shared scale. They cannot be added, they cannot be ranked, and a trade-off between two of them is not a meaningful operation.

1.2 What this work replaced

A common enterprise measurement architecture allocates budget using a multi-touch attribution model: a large, heavily parameterized system with a long chain of transformations and channel and touchpoint weights that can be difficult to verify independently. Model sophistication alone does not solve the governance problem. When a channel's credit shifts, it may be unclear whether business performance changed or whether the model's assumptions, data coverage, or weights changed.

In these environments, an independent measurement layer can improve neutrality without fully resolving reproducibility. Section 3 explains why independence and reproducibility should be treated as separate requirements.

1.3 Contribution and scope

The protocol described here replaces attribution-based allocation with a retailer-owned randomized comparison. Eligible customers are assigned through a first-party identity graph into a persistent portfolio holdout and an addressable media population, paid delivery is suppressed to the holdout, and the difference is read on total omnichannel sales.

The contributions are specific:

1. A persistent randomized portfolio counterfactual defined on the retailer's first-party identity spine, with outcomes read from the retailer's own omnichannel transaction ledger.
2. A protocol-level distinction between the portfolio estimand and channel-specific estimands: the global holdout identifies the effect of the addressable media portfolio, while a causal channel comparison requires a separate randomized channel holdout.
3. Common rules for assignment, outcome definition, read windows, confidence intervals, and compliance diagnostics so that platform-specific attribution choices do not determine the causal result.
4. A division of channels into person-level, geo, and out-of-protocol tiers, with suppression leakage, match coverage, and other limitations reported explicitly rather than absorbed into an opaque adjustment.

A note on scope. The protocol does not attempt to harmonize platform-reported lift. It standardizes the retailer-side causal design and outcome source. The always-on portfolio holdout establishes the total effect of addressable media; it is not a decomposition of that effect by channel. Where channel-level causal comparisons are required, they must come from separate randomization under the same protocol.

What this paper does not do. It reports no quantitative results. The protocol is presented as an operational framework rather than an evaluation of a named retailer or deployment. Accordingly, it does not report portfolio lift magnitudes, incremental return on ad spend (iROAS) values, leakage rates, or before-and-after business comparisons. The paper should be read as a methods and architecture contribution.

Section 2 positions the work. Section 3 develops the two-layer problem that motivates it. Section 4 explains why platform-reported figures are not directly comparable. Sections 5 to 7 describe the randomized protocol, its data requirements, and the practical limits of network support. Section 8 explains why platform lift is excluded from the causal path. Section 9 covers the geo tier. Section 10 addresses suppression leakage and conservative interpretation. Section 11 states the estimand boundaries. Section 12 addresses incentives and measurement governance. Sections 13 to 15 cover evidence, limitations, and future work, and Section 16 concludes.

2. Background and Related Work

The economics of measuring advertising returns. The difficulty of detecting advertising effects with observational data is well established. Lewis and Rao [1] demonstrate that the statistical power required to measure returns to advertising is far greater than practitioners typically assume, and that the effect sizes of interest are small relative to the variance of individual purchase behavior. This result underpins the entire design here: if effects are hard to detect even with experiments, then estimates produced without one deserve considerably less confidence than they usually receive.

Experimental versus observational estimates. Blake, Nosko and Tadelis [2] show through large-scale field experimentation that paid search returns were substantially overstated by conventional attribution, particularly for branded terms where the counterfactual purchase would largely have occurred anyway. Gordon, Zettelmeyer, Bhargava and Chapsky [3] compare observational approaches against large field experiments and find that commonly used observational methods do not reliably recover experimental estimates, with the

direction and magnitude of the discrepancy varying by campaign. Together, these studies establish the empirical case for prioritizing randomized evidence in the use case considered here. Observational causal methods may provide credible estimates under appropriate assumptions, but randomized assignment offers the preferred counterfactual for this protocol.

Bridging attribution and causal measurement. Recent practitioner-facing work has explored predictive approaches that attempt to connect attribution outputs to causal estimates [4]. That direction is complementary to this one but differs in an important respect: it seeks to improve the correspondence between a modeled attribution figure and a causal quantity, whereas the position taken here is that the modeled figure should be removed from the decision path rather than corrected.

The contribution is therefore not a new experimental estimator. It is an operating framework for applying randomized measurement consistently across networks with different data structures, identity systems, and measurement conventions. The framework addresses the architecture and governance required to make experimental measurement a continuous organizational capability rather than an occasional study.

3. The Two-Layer Measurement Problem

This section describes a recurring measurement limitation that can persist even after an independent measurement layer is introduced.

3.1 Independence Alone Is Insufficient

In one common implementation, an independent measurement layer is introduced so that an attribution model does not rely directly on platform self-reporting.

This layer can improve neutrality without making the resulting estimate independently reproducible. Addressing the most visible conflict-of-interest concern may increase confidence even when the underlying methodology remains difficult to inspect or reproduce.

3.2 Modeled output inside a modeled system

The structural issue arises when modeled output feeds another modeled system. An independent measurement layer may make choices about window definitions, exposure criteria, counterfactual construction, identity resolution, and normalization across sources within a proprietary methodology that can be described at a high level but not independently reproduced. An attribution model may then use that output while adding its own path reconstruction, touchpoint classification, decay, and weighting.

Neither layer was fully reproducible from outside its own methodology. Combined, they made it difficult to test which assumptions were driving a change in the final result. Independence and transparency are distinct properties; solving the first does not automatically provide the second.

3.3 Diagnostic Limitations of the Prior Architecture

A central diagnostic limitation is the inability to decompose a change. When a channel's contribution moves, several explanations may be plausible: business performance changed, the independent measurement methodology or data coverage shifted, or the attribution model's weights changed during refresh. Without diagnostics that separate these causes, confidence in the resulting estimate can decline.

Channel-level attribution totals may also fail to reconcile consistently with observed transaction totals. When independent reads are periodic but the attribution model runs continuously, decisions between calibration points depend on model extrapolation rather than a contemporaneous experimental read.

A further constraint is the quality and granularity of the underlying inputs. Independent measurement layers may receive windowed, aggregated, or partly modeled conversion feeds and face the same identity-resolution limits as the source platforms. Independent interpretation cannot recover information that is absent from the source data. Cross-client benchmarks also require caution because omnichannel retailers with substantial store-based revenue and heterogeneous customer segments may differ materially from benchmark populations.

3.4 Simplifying the Causal Measurement Architecture

The problem is therefore not simply whose model is used. Large decisions can still depend on stacked estimates that cannot be reproduced end to end, regardless of the credibility of the parties producing them.

The replacement framework reduces the number of modeled layers in the causal path. Random assignment is defined internally, suppression rules are documented, outcomes come from one omnichannel sales ledger, and

the causal contrast is the difference between randomized populations. Platform data is used to describe delivery and compliance rather than to determine the conversion outcome.

4. Why Cross-Network Figures Are Not Comparable

Four independent sources of incomparability operate simultaneously, and any one of them is sufficient to prevent two figures from being traded off.

4.1 Four counterfactuals, one word

The most consequential difference across platform lift estimates is often the definition of the counterfactual. Table 1 summarizes several constructions that may all be reported under the label "lift."

Table 1. Counterfactual constructions reported under the same label.

Construction	What the comparison is against	Observable externally
Placebo control	People shown neutral creative; depending on implementation, the study can alter auction or optimization behavior	No
Ghost bids	Auctions the network would have won but intentionally did not serve; a strong within-network counterfactual but not externally observable	Only inside that network
Managed lift study	A randomized cell constructed by the network under its own study design	No
No baseline stated	Attributed conversions, implicitly baselined at zero, treating every attributed purchase as caused	Not applicable

These constructions operate on different scales. Directly averaging or ranking estimates produced under different counterfactuals generally lacks a common causal interpretation, even when each is reported under the same label.

4.2 Conversion definitions

On conversions the disagreements are numerous enough that enumerating them cleanly is difficult. Table 2 records the ones encountered most often.

Table 2. Ways two networks can both report "a conversion" and mean different things.

Dimension of disagreement	One convention	Another convention
Qualifying interaction	View-through with a multi-day lookback	Click-through inside a short window
Timestamp assignment	Stamped at the click	Stamped at the purchase
Repeat handling	Every conversion in a session counted	Deduplicated to one per user per window
Observation status	Observed conversions only	Modeled conversions blended in, not separated
Campaign scope	Once per user	Once per campaign, so three campaigns yield three
The unit itself	An order	A transaction line, a basket, a subscription start, or an inferred store visit

The timestamp row is the one that causes the most confusion in practice, because it moves the same order between reporting periods without changing anything about the order, and two systems disagreeing only on that dimension will still disagree on every periodic total.

4.3 Identity units

Identity definitions are also fractured, and the units are not interchangeable. A network may report activity at the person, device, household, or probabilistic-cluster level; Table 3 summarizes the implications for comparison and suppression design.

Table 3. Identity units, and what each one is.

Unit	What it represents	Consequence for comparison
Logged-in person	An authenticated individual	Preferred unit for person-level randomization, suppression, and retailer outcome joins
Device	A browser or application instance	One person may hold several; several people may share one
Household	A residence or connected-television account	Coarser than a person-level identity spine; requires an explicit household assignment or roll-up rule
Probabilistic cluster	An inferred grouping	Accuracy unstated and not externally checkable

A lift figure computed on households is not comparable to one computed on individuals before any other difference is considered, and a figure computed on a probabilistic cluster carries an unknown amount of the cluster's own error.

4.4 Exposure definitions and reporting clocks

Exposure definitions diverge on whether an unviewable impression counts, whether a sub-second video view counts, whether a scroll-past counts, and how reach and frequency are constructed. These differences matter for compliance diagnostics and exposure-based secondary analyses. The primary randomized estimate avoids conditioning on a network's exposure definition, but inconsistent exposure logs still make it harder to verify treatment separation across channels.

Reporting clocks also differ. Exposure logs can arrive late, use different day boundaries, or be restated after an initial read. Because the primary outcome comes from the retailer's own transaction ledger, those network timing differences do not redefine the conversion outcome; they can, however, change compliance diagnostics and exposure-based sensitivity analyses if they are not versioned consistently.

4.5 Why attribution compounds rather than resolves

Conventional attribution does not fix any of this. It allocates credit for purchases that already occurred based on observable path position, which systematically rewards whatever sits nearest the conversion and penalizes upper-funnel exposure that leaves no click trail. Its weights depend on observable paths, and that signal continues to degrade, so credit drifts toward lower-funnel channels and shifts between refreshes for reasons that cannot be verified from outside the model.

5. The Protocol

5.1 The randomized counterfactual

The causal counterfactual is created by assignment, not predicted from customer history. Eligible customers are assigned by stable hash to a persistent portfolio holdout or an addressable media population before outcomes are observed. The holdout is suppressed from addressable paid media as completely as network capabilities allow. The primary estimate is the intent-to-treat difference in total omnichannel sales between the randomized populations, normalized for population size and read over a pre-specified window.

Pre-period customer behavior may be used for balance checks, stratification, or a pre-specified regression adjustment to improve precision, but it does not determine how much of an individual order is credited to media. A one-hundred-dollar order is not split into modeled "organic" and "media-caused" portions; causality is inferred from the randomized difference across populations.

This distinction keeps the causal claim tied to randomization. Network attribution logic, customer propensity scores, and path weights may be useful diagnostics, but they are not the source of the counterfactual.

5.2 Protocol-level constants and compliance diagnostics

Three elements are standardized across reads: assignment rules, the omnichannel outcome definition, and the lag/read-window policy. Match rate, suppression compliance, and exposure timing are reported as diagnostics because they affect interpretation, power, and the degree of treatment separation achieved in practice.

Match rate is not treated as an automatic causal correction. The primary intent-to-treat estimate remains defined by randomized assignment. If a treatment-on-the-treated estimate is required, assignment can be used as an instrument for exposure only under explicit compliance assumptions and with the additional uncertainty reported.

A common lag and read-window policy replaces network-specific attribution windows. Timestamped exposure is used to diagnose delivery, contamination, and timing and to support sensitivity analysis; it is not transformed into a substitute counterfactual.

These choices reduce the number of platform-specific assumptions entering the causal estimate while keeping known compliance gaps visible.

5.3 A compliant measurement end to end

Table 4 sets out the ordered protocol from assignment through reporting.

Table 4. The protocol, start to finish.

Step	Action	Why it is placed here
1	Randomize the eligible first-party audience into a persistent portfolio holdout and addressable media population using a stable assignment	Creates the causal counterfactual before outcomes are observed
2	Pre-register the primary outcome, estimand, minimum detectable effect, holdout share, and read window	Prevents specification changes after results are visible
3	Push suppression to each network and record match coverage, implementation status, and expected compliance evidence	Makes treatment delivery and known limitations auditable
4	Run the holdout while collecting spend, delivery, reach, and leakage/compliance diagnostics without changing assignment	Preserves the randomized treatment definition during the read
5	Join both randomized populations to the retailer's internal omnichannel sales ledger	Keeps the causal outcome independent of platform attribution rules
6	Estimate the intent-to-treat difference in omnichannel sales; apply only pre-specified covariate adjustment for precision	Randomization determines causality; pre-period covariates may improve precision without replacing the counterfactual
7	When a channel-specific causal effect is required, run a separate randomized channel holdout under the same outcome and window rules	The global portfolio holdout does not identify additive channel contributions
8	Report effect size, incremental revenue, applicable iROAS, confidence interval, and match/leakage diagnostics	Separates the causal estimate from the evidence about implementation quality

The protocol deliberately separates causal estimation from delivery diagnostics. Randomized assignment determines the causal contrast; platform exposure and match metrics describe compliance. Conditioning the primary read on observed exposure can reintroduce selection bias and is therefore reserved for secondary analyses with explicit assumptions.

5.4 Why the portfolio holdout is persistent

The portfolio control population is held out continuously rather than reconstructed for each campaign. That persistence creates a standing counterfactual that can be read whenever a business decision requires it, subject to pre-specified reporting windows.

The persistent global holdout answers a portfolio question: what happens to total omnichannel sales when addressable paid media is removed for a randomized customer population while the rest of the commercial environment proceeds as observed? This is the causal effect identified by the always-on design.

It does not identify additive channel contributions. A channel-specific holdout, in which only one channel is suppressed while other media continues, estimates the marginal effect of that channel conditional on the current media mix. That can be useful for allocation, but it is a different estimand. If channels are to be compared, those experiments should use the same outcome definition and read-window rules and should be run concurrently, or with a design that accounts for changing background conditions. Channel estimates should not be forced to sum to the global portfolio effect because cross-channel interactions and overlap can make additivity fail.

The commercial cost of persistence should be stated explicitly. A permanently suppressed population may forgo incremental revenue that the media would otherwise have generated. The holdout share is therefore both a statistical and business decision: too small and the minimum detectable effect becomes unhelpful; too large and the opportunity cost of measurement becomes material. The share should be pre-registered rather than adjusted after results are observed.

Persistence also creates a failure mode that short campaign tests reduce: cohort drift. Over long periods, attrition, life-stage changes, identity resolution, or other non-media differences can cause the randomized populations to diverge. The cohorts should therefore be monitored on pre-treatment and non-media characteristics, with a documented refresh rule if comparability deteriorates.

6. What the Protocol Requires From a Network

The engineering challenge is not the number of required fields but obtaining them consistently at the grain needed to monitor treatment delivery and suppression.

The preferred inputs are: a stable hashed identifier that maps to the retailer's first-party identity spine; enforceable suppression of the randomized holdout; confirmation of audience match coverage; timestamped exposure or reach information where privacy rules permit; spend aligned to the same campaign and time grain; and enough delivery information to estimate or bound control-group leakage. The causal outcome itself remains on the retailer's internal sales ledger.

Platform conversion data is deliberately excluded from the primary causal estimate. Networks provide delivery and cost information, while orders, baskets, margin, and store transactions come from the internal ledger joined to the randomized identity spine. Accepting a platform conversion count would also import that platform's attribution window, credit rule, modeling choices, and outcome coverage, undermining the common retailer-side outcome definition.

The required network feed can therefore be summarized as who was eligible, who was reached, when delivery occurred, and what it cost.

7. What Networks Actually Expose

Network capabilities vary materially by channel class; Table 5 summarizes the resulting protocol tiers.

Table 5. Channel classes against protocol requirements.

Channel class	Exposure grain available	Suppression	Protocol tier
Walled-garden search and social	Impressions, clicks, spend, reach, frequency at campaign or ad-set aggregate; person grain only inside a clean room	Supported, enforcement rarely verifiable	Person-level, with leakage

Channel class	Exposure grain available	Suppression	Protocol tier
Programmatic demand-side platforms	Log-level impressions with a persistent identity key; closest to native requirement	Generally supported	Person-level
Addressable connected television	Household level rather than person level	Variable	Household-level or geo, depending on assignment design
Off-site retail media	Aggregate delivery; conversions attributed against the host retailer's transactions	Variable	Outside the strict protocol
Linear television, out-of-home, in-store media	No identity layer; resolves only to a market	Not applicable	Geo tier

Several structural gaps recur across these classes.

The unexposed record is missing. Networks report who they reached. They almost never report who they could have reached and chose not to. Without that ghost-bid population, a within-network control cannot be constructed from their data at all, only from an externally supplied suppression list.

Suppression cannot always be verified. An exclusion audience can be submitted, but delivery against it may not be observable at person level. This creates noncompliance and possible control contamination. Under a standard positive-treatment interpretation, contamination will often attenuate the intent-to-treat difference toward zero, but the exact direction and magnitude should not be treated as known without additional assumptions.

Audience match rates are often reported with limited methodological transparency. A hashed audience may be uploaded and a rounded figure returned, sometimes subject to privacy thresholds and without clarity on how unmatched records are handled. In this protocol, match rate is primarily a compliance and coverage diagnostic; it becomes part of a causal adjustment only when a treatment-on-the-treated analysis is explicitly justified.

Clean-room asymmetry further limits standardization. Person-level joins may be available only inside a network-specific clean room with its own query restrictions, aggregation thresholds, and export limits. As a result, identical compliance analyses may not be executable at the same granularity across networks, especially for small holdout cells and segment-level cuts.

Exposure timing can also vary across sources. Delivery logs may arrive on different clocks or be restated after initial reporting, which affects compliance diagnostics and exposure-based sensitivity analyses. The primary transaction outcome remains anchored to the retailer's own reporting clock and pre-specified read window.

8. Excluding Platform-Reported Lift from the Primary Estimand

The protocol does not reconcile platform lift estimates; it keeps them outside the primary causal calculation. Directly averaging lift estimates produced under different counterfactuals, identity definitions, outcome rules, and windows generally lacks a common causal interpretation. The protocol instead uses a common randomized counterfactual and retailer-owned outcome.

For the portfolio read, networks are treatment-delivery and exposure sources rather than measurement authorities. A network may still run its own lift study, but that result is treated as supplementary evidence rather than an input to the retailer-side causal estimate. A network-specific causal claim is made only when a separate randomized channel design supports it.

The consequences are straightforward. Network-declared attribution windows do not determine the primary read; the protocol uses one pre-specified outcome window. Differential match rates are disclosed as compliance

information rather than assumed to be fully neutralized by a correction factor. Exposure data supports diagnostics, but the causal contrast remains based on randomized assignment.

A platform cannot directly inflate the primary causal estimate through its own conversion model because platform-reported conversions and lift do not enter the calculation. That does not eliminate all bias: suppression leakage, identity fragmentation, spillovers, implementation errors, and selective reporting of delivery data remain important risks.

Where person-level suppression or compliance monitoring is not feasible, the channel moves to the geo tier and is measured with a coarser experimental design under a different set of assumptions rather than with a modeled substitute presented as equivalent.

8.1 Trade-offs of Excluding Platform-Specific Measurement

Declining to ingest network-reported measurement is not free, and the costs should be set against the benefit rather than assumed away.

The protocol forfeits information that networks genuinely hold and the retailer does not. A network running ghost-bid logic has access to a counterfactual that is, on its own terms, cleaner than an externally imposed suppression list, because it compares against auctions it would have won rather than against a population it was asked not to reach. That construction is unavailable from outside, and refusing the network's read means refusing that information along with everything else.

It also forfeits within-network granularity. A network can decompose its own delivery in ways the retailer cannot reconstruct from an exposure feed, and analyses that would benefit from that decomposition are not available under this design.

The protocol therefore trades some platform-specific detail for reproducibility and auditability. A retailer operating at a different scale, with stronger access to network-native experiments or a different tolerance for unverifiable inputs, may reasonably make a different trade-off.

9. The Non-Addressable Tier

A meaningful share of media has no identity layer at all. Linear television, out-of-home, and in-store media resolve only to a market, so no protocol requirement can be satisfied for them. This is why geo testing is a permanent component of the design rather than a fallback.

Geo testing serves two functions. For non-addressable inventory it may be the only practical experimental option. For addressable inventory it provides an independent validation mechanism that does not rely on individual-level identity matching or person-level suppression. Geo tests are not assumption-free: market comparability, spillover, local shocks, and statistical power must still be managed. Periodic paired reads are therefore best treated as external validation rather than as a perfect bound.

Addressable connected television represents an intermediate case because delivery is commonly household-level while the retailer's identity spine may be person-level. A defensible implementation must either randomize and analyze at a consistent household unit, define a documented household roll-up, or move the inventory to a coarser tier. The unit conversion is an assumption and should be stated explicitly.

Off-site retail media is outside the strict protocol by construction. Exposure and spend can be captured, but outcomes land on another retailer's transactions and cannot be joined to the internal customer ledger on a shared basis. No amount of protocol design resolves this, because the outcome data belongs to a different party measuring a different ledger.

10. Suppression Leakage and Conservative Interpretation

Suppression is not airtight, and no claim is made that it is. Control-group exposure and treatment-group non-delivery create noncompliance. Under common assumptions in which media effects are non-negative and leakage mainly reduces treatment separation, the intent-to-treat estimate will tend to move toward zero. That makes a conservative interpretation plausible, but not a guaranteed mathematical lower bound in every setting.

Results should therefore be reported as causal effect estimates with confidence intervals together with match coverage, suppression leakage, and other compliance diagnostics. Where the required assumptions are credible, management may choose to interpret the estimate conservatively; the methodology should not present attenuation as certain when the implementation cannot establish it.

Periodic geo experiments provide a useful independent check because they do not use person-level match rates or individual suppression. Agreement across the two designs increases confidence; disagreement is a diagnostic that the identity, suppression, geographic design, or estimand definitions need investigation.
The current limitation is that this validation is periodic rather than continuous, and the protocol does not yet produce a formal per-channel compliance-adjusted effect with identified bounds. Section 15 identifies this as a priority for future work.

11. What the Protocol Does Not Measure

A measurement system is easier to trust when the boundary of its estimand is explicit. Three important quantities sit outside the primary portfolio read.
Cannibalization is not separated. The protocol reads the gap on total omnichannel sales, so media-driven demand that displaces demand the business would have captured through another route is netted out at the aggregate level. That is appropriate for a portfolio-level sales estimand but insufficient for a category- or product-level question. A campaign may materially change product mix while producing little change in total sales. Category-level experimental reads can address that question, at the cost of lower statistical power as the outcome is segmented more finely.
Long-run effects are truncated by the read window. The window is held open long enough to cover the purchase cycle, which captures the response the campaign was intended to produce. It does not capture effects that accumulate over periods longer than that window, whether brand familiarity, changes in category consideration, or shifts in purchase frequency that develop over years. A persistent holdout is in principle the right instrument for measuring these, because the same customers remain unexposed for extended periods, but doing so requires reading the same cohorts over a much longer horizon than allocation decisions operate on, and the cohort drift discussed in Section 5.4 becomes the binding constraint on how long that can run.
Competitive response is not manipulated by the experiment. Randomization occurs while competitors continue to advertise and market conditions evolve as observed. The estimand is therefore the effect of removing this retailer's addressable media under the prevailing competitive environment, not the effect of changing total category advertising pressure. First-party randomization does not identify the latter.
None of these is a defect in the design so much as a statement of what a subtraction on total sales can and cannot recover. They are recorded here because a measurement system that is silent about its boundary invites its outputs to be used past it, and the failure described in Section 3 was in part a failure of exactly that kind.

12. Incentives and Measurement Governance

Removing platform-reported lift from the causal path does not eliminate all sources of measurement risk. It shifts the remaining risks toward implementation, observability, and participation, which should be monitored explicitly.

12.1 Outcome measurement and self-reporting

Keeping outcomes on the retailer side removes platform-reported conversions from the primary causal calculation. A network may still report its own lift for supplementary use, but the primary estimate does not depend on that figure. The remaining risks therefore concern delivery, identity, and compliance information that the retailer cannot observe independently.

12.2 Exposure, Suppression, and Participation Constraints

A network can influence the quality of compliance evidence through the exposure records, match-rate information, and suppression reporting it makes available. Reach-reporting error can distort apparent delivery efficiency, while unobserved suppression leakage can reduce separation between treatment and control groups. Data-sharing incentives may not always align with independent measurement requirements, particularly when participation requires additional suppression, identity-resolution, or delivery reporting. Coverage therefore depends partly on commercial arrangements and data-access agreements, not only on methodology.
Table 6 summarizes where each incentive surfaces and how well the design answers it.

Table 6. Residual measurement risks under the protocol.

Vector	What the network controls	Detection	Residual risk
Platform-reported lift	The platform's own conversion and lift methodology	Exclude from the primary causal calculation; compare only as supplementary evidence	Cannot directly change the retailer-side causal estimate
Reach-reporting error	Claimed delivery records	Audit against available impression logs, clean-room delivery, and expected reach	Low to moderate, depending on observability
Unobserved suppression leakage	Delivery against the excluded audience	Monitor excluded-audience delivery where possible and compare against match coverage	Moderate when person-level verification is unavailable
Match-rate error	Reported audience resolution	Cross-check against delivered reach and clean-room or identity-partner evidence where available	Moderate; reported figures may be rounded or privacy-thresholded
Declining participation	Whether to support suppression and provide delivery evidence	Visible as a coverage limitation	Not a statistical bias in measured channels, but a protocol-coverage risk

12.3 Why the Governance Problem Cannot Be Fully Eliminated

No measurement design eliminates every implementation risk; it determines which quantities remain open to error or limited observability. Under self-reported measurement, the platform's outcome definition directly affects the reported result. Under this protocol, the remaining uncertainty is concentrated in delivery, identity, suppression evidence, and willingness to participate. These risks are further from the outcome calculation but are not eliminated.

The practical implication is that the protocol must be operated continuously rather than merely documented. Suppression audits, match-coverage checks, balance monitoring, and data-quality reviews protect the estimate only when they are performed consistently. A control framework that depends on recurring verification can degrade if those checks lapse.

This is one reason to keep as much of the design as possible on the retailer's own side of the boundary. Assignment and outcomes can be audited internally; delivery and identity evidence that remains inside a network can only be validated to the extent the network permits.

13. Evidence and Its Limits

No quantitative results are reported in this paper. The protocol is specified as an operating architecture for addressable digital media rather than evaluated as a named deployment. It supports the computation of portfolio incrementality, incremental revenue, and iROAS from randomized reads. Because no deployment data are reported, the paper's claims are limited to the design, assumptions, and governance properties of the framework.

This limits what the paper can establish, and Table 7 states the evidence status explicitly.

Table 7. Evidence status by claim class.

Claim class	Status
Protocol specified as an operating architecture across addressable digital channels	Operational framework described; no named deployment evaluated

Claim class	Status
Portfolio outputs can support budget decision-making under common rules	Specified as an intended decision use; no organizational adoption claim made
Portfolio iROAS, incremental revenue, and incrementality can be computed	Defined by the methodology; no empirical values reported
Channel-level causal rankings from the global holdout alone	Not identified by the portfolio design; requires separate channel randomization
Empirical comparison of business figures before and after adoption	Not evaluated
Suppression leakage and per-channel noncompliance	Specified as a diagnostic; no empirical rates reported
Geo-versus-audience validation agreement	Specified as periodic validation; no quantitative agreement reported
Comparison with attribution-based allocation	Supported conceptually by reproducibility and causal design; not demonstrated numerically

The last rows deserve emphasis. The paper's defensible claim is that the portfolio estimate can be reproduced from a documented randomization, outcome source, and reporting protocol. It does not demonstrate numerically that the estimate is closer to truth than every alternative, and it does not claim that the global holdout alone identifies causal channel rankings. Those stronger claims require external validation or separate channel-level randomization.

14. Limitations

The portfolio holdout does not identify channel contributions. This is the most important scope limitation. The global randomized difference identifies the incremental effect of the addressable media portfolio as a whole. Channel-level causal effects require separate randomization and should not be inferred by decomposing the portfolio lift with attribution weights.

Suppression noncompliance can attenuate treatment separation. A network that cannot reliably exclude the randomized holdout, or cannot provide enough evidence to assess leakage, weakens the intent-to-treat contrast. Person-level exposure is therefore highly valuable for compliance monitoring, even though the causal estimand itself is defined by assignment.

Small channel-specific experiments are statistically difficult. Minimum detectable effect depends on volume and outcome variance, and noncompliance can further reduce effective treatment separation. A low-spend tactic can therefore produce an inconclusive estimate even when its true effect is non-zero.

Privacy constraints limit compliance diagnostics. Person-level joins increasingly occur only inside clean rooms with query limits, aggregation floors, and export restrictions. Those differences make identical auditing difficult across networks and can leave match or leakage measures available only as ranges.

Randomization is only as reliable as the identity graph on which it is implemented. Fragmented identifiers can place the same customer into inconsistent delivery states across channels, while customers absent from the addressable file cannot be assigned. The identity spine is therefore both the sampling frame and a potential source of treatment contamination.

Persistent cohorts can drift over time. Attrition, identity changes, or other non-media differences may gradually reduce comparability between the randomized populations. The protocol requires balance monitoring and a documented refresh rule, but the optimal refresh cadence is not established here.

Household-level addressable television requires an explicit unit-of-analysis decision. A person-level identity graph cannot be assumed to map one-to-one to a household delivery unit, so the reconciliation must be documented or the inventory measured at a coarser level.

No empirical outcome magnitudes are reported. As Section 13 explains, readers can evaluate the architecture and assumptions but cannot independently assess the size, stability, or business impact of measured effects in a specific deployment.

15. Future Work

The list below is ordered by how much each item would improve confidence in future applications of the framework rather than by how interesting it is.

Formalize noncompliance estimation. Estimating assignment-to-exposure compliance and developing transparent bounds, or an instrumental-variable treatment-on-the-treated estimate where assumptions are defensible, would quantify the effect of suppression leakage rather than describing it only qualitatively.

Standardize concurrent channel-specific holdouts. If the protocol is to support direct channel allocation, each causal channel estimate should come from a randomized design using the same retailer-owned outcome, reporting window, and compliance standards. The global holdout can then serve as a portfolio anchor rather than an additive decomposition target.

Read lift by category as well as in total. Section 11 notes that cannibalization is netted out in the total-sales estimand. Category-level randomized reads would begin to expose mix effects, with statistical power determining how far the outcome can be segmented.

Build an incentive-compatible partner tier. Preferential measurement or budget consideration for networks that provide stronger suppression verification and delivery transparency could align commercial incentives with protocol participation, subject to normal procurement and governance controls.

Define cohort-refresh and interference checks. Future work should specify thresholds for persistent-cohort drift, cross-device or household spillover, and identity fragmentation so that an always-on holdout can remain persistent without allowing its randomization assumptions to degrade unnoticed.

16. Conclusion

The cross-network measurement problem is not simply that one platform's number is right and another's is wrong. The numbers are produced under different counterfactuals, identity rules, attribution windows, and outcome definitions, so they should not be treated as interchangeable causal estimates.

This paper describes an alternative that moves the causal counterfactual to the retailer side. A persistent randomized portfolio holdout is defined on the first-party identity spine, networks provide delivery and cost information, and outcomes come from the retailer's omnichannel transaction ledger. The resulting intent-to-treat difference identifies portfolio-level incrementality. It does not, by itself, identify additive channel contributions; those require separate channel-level randomization under the same protocol.

The design accepts practical limitations. Suppression can leak, identity resolution is imperfect, and privacy restrictions can limit compliance monitoring. Results should therefore be reported with confidence intervals and explicit leakage and match diagnostics rather than as precision that the implementation cannot support. Periodic geo experiments provide an independent validation path with their own assumptions.

The paper does not report empirical outcomes, so its contribution is methodological rather than a numerical demonstration of superiority. The next steps are to quantify noncompliance, standardize channel-specific randomized reads where allocation requires them, and define refresh rules that preserve the validity of a persistent holdout over time.

References

[1] R. A. Lewis and J. M. Rao, "The Unfavorable Economics of Measuring the Returns to Advertising," The Quarterly Journal of Economics, vol. 130, no. 4, pp. 1941-1973, 2015. doi: 10.1093/qje/qjv023.

[2] T. Blake, C. Nosko and S. Tadelis, "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment," Econometrica, vol. 83, no. 1, pp. 155-174, 2015. doi: 10.3982/ECTA12423.
[3] B. R. Gordon, F. Zettelmeyer, N. Bhargava and D. Chapsky, "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," Marketing Science, vol. 38, no. 2, pp. 193-225, 2019. doi: 10.1287/mksc.2018.1135.
[4] S. Oh, "Predictive incrementality: Bridging attribution and causal measurement for real return on marketing investment," Journal of Digital & Social Media Marketing, vol. 13, no. 4, p. 310, 2026. doi: 10.69554/wxoo3121.