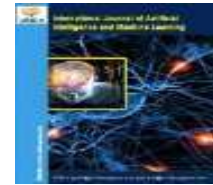




International Journal of Artificial Intelligence and Machine Learning

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



Research Paper

Open Access

Testing Paid Media at Speed and Scale: A Decision Framework for Omnichannel Retailers

Venkata Suneel Dasika

Independent Researcher, India. E-mail: dasikasuneel@gmail.com

Abstract

U.S. retail media advertising spend is projected to reach \$69.3 billion by 2026, yet large omnichannel retailers lack a structured framework for determining whether that spend drives real incremental value. Four structural forces make causal paid media measurement uniquely difficult: (1) the offline conversion gap, in which the majority of purchase outcomes are invisible to digital attribution systems; (2) the walled garden self-attribution problem, where competing platforms simultaneously claim credit for the same conversion; (3) accelerating user-level signal loss driven by privacy regulation; and (4) a first-party data paradox in which enterprise CRM richness is undermined by the inability to construct clean control groups when multiple teams draw from the same customer pool. This paper introduces a two-tier decision framework addressing these challenges. Tier 1 presents high-confidence causal methods (geo market testing, global customer holdout, difference-in-differences / synthetic control, and randomized platform A/B with cross-channel exclusion). Tier 2 presents speed-oriented directional methods (CUPED-adjusted pre/post, switchback testing, standard platform lift studies, and short-window marketing mix modeling). For each, the paper examines mechanics, omnichannel applicability, and structured pros and cons.

Keywords: paid media measurement, omnichannel retail, incrementality testing, geo testing, marketing mix modeling, causal inference, A/B testing, customer holdout, difference-in-differences

I. Introduction

U.S. retail media advertising spend crossed \$53.7 billion in 2024 and is projected to reach \$69.3 billion by 2026 — growing faster than Meta and Alphabet combined [1]. Yet for large omnichannel retailers, a foundational question remains stubbornly difficult to answer: did the campaign actually work? Last-touch, multi-touch, and marketing mix models each answer this differently, often by wide margins, and none produces a causal estimate of incremental lift [2].

The challenge is acute because an estimated 80% of retail transactions still complete in physical stores [3], even when the customer's path began with a digital ad. At the same time, marketing teams face real organizational constraints — campaign flights close, competitive pressures mount, and budget decisions must be made quarterly. Practitioners face a genuine tradeoff between causal confidence and speed. This paper resolves that tradeoff with a tiered decision framework matching the right methodology to the right measurement context.

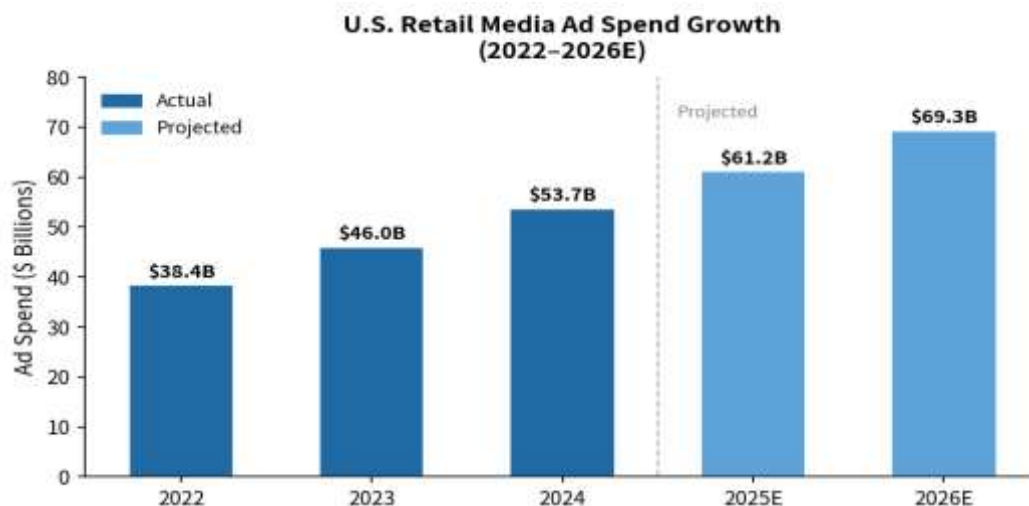


Figure 1. U.S. Retail Media Ad Spend, 2022–2026E (eMarketer, 2025)

2. The Omnichannel Measurement Landscape

Four distinct forces make causal inference in large omnichannel retail structurally more difficult than in pure-play digital environments. Together, they create conditions in which no passive measurement system can reliably answer the question a marketer most needs answered: would this sale have occurred without the media spend?

2.1 Offline Conversion Gap

For large-format retailers, the majority of purchase outcomes are invisible to digital measurement. Research suggests 84–85% of the true sales impact from online advertising manifests offline [6] — the dominant outcome for home improvement, appliances, and building materials. Standard attribution tools have no mechanism to observe store-completed purchases, producing a systematic underestimation of upper-funnel channels such as CTV and display.

2.2 Walled Garden Self-Attribution

Each platform reports performance using its own attribution logic, producing systematic double-counting: a single purchase is routinely claimed by multiple platforms simultaneously, inflating portfolio ROAS while true incremental revenue remains unknown. The only methodology that resolves this is a controlled experiment where a portion of the audience is deliberately withheld from exposure.

2.3 Privacy-Era Signal Loss

Apple's ATT framework reduced observable conversion rates on iOS by 30–60% for many advertisers [7]. Third-party cookie deprecation and GDPR/CCPA expansion further erode cross-device attribution. Paradoxically, this increases the value of active experimentation — particularly geo-level and aggregate methods that derive causal inference from experimental design rather than user-level tracking.

2.4 The First-Party Data Paradox

Large retailers possess enormous CRM assets — tens of millions of identified customers with purchase histories and category affinities. Yet this richness creates a structural problem: when multiple teams (paid media, email, loyalty, CRM, direct mail) draw concurrently from the same customer pool, clean control groups become organizationally elusive. A paid social holdout group simultaneously receives email promotions, loyalty pushes, and CRM offers from the same brand — it is a "no-paid-social" group, not a true no-treatment group. This contaminates experiments and systematically understates measured incrementality.

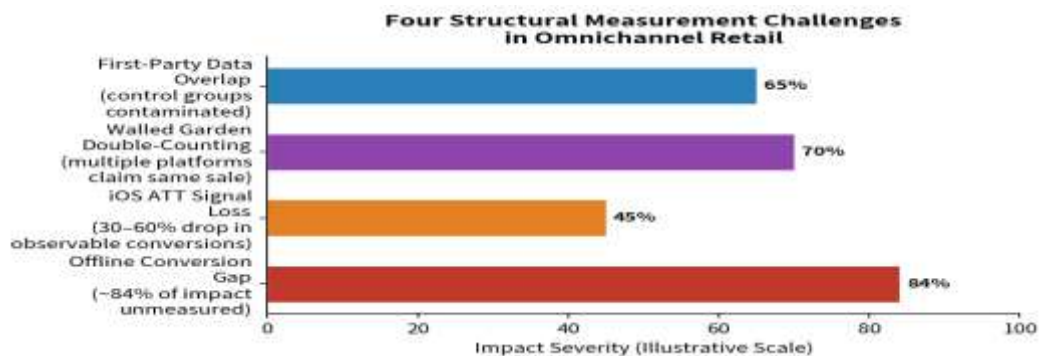


Figure 2. Relative impact of four structural measurement challenges in omnichannel retail

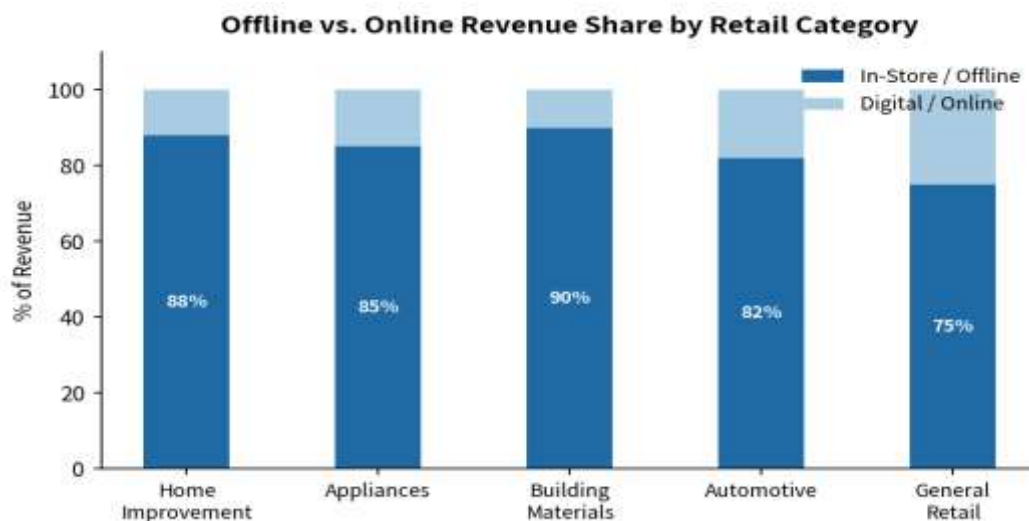


Figure 3. Offline vs. Online Revenue Share by Retail Category (illustrative, industry estimates)

3. Tier 1: High-Confidence Causal Methods

Tier 1 methods are appropriate when the measurement question justifies weeks of runtime, meaningful budget holdouts, and dedicated analytical resources. Causal validity is built into the experimental structure — not inferred statistically after the fact.

3.1 Geo Market Testing

The unit of randomization is a geographic entity (DMA, ZIP cluster, store trade zone) rather than an individual user. A subset of markets receives full media weight; matched markets are held out entirely. The difference in total sales — online and in-store — across treatment and control markets constitutes incremental lift. Pre-period revenue trend (70–80% of matching weight) is the dominant market-matching variable; structural covariates (homeownership, store density, media cost index, seasonality) supplement it. Treatment-market spend should represent 0.5–1.5% of baseline revenue to clear natural variance. Runtime: 4–8 weeks minimum. The customer overlap problem is substantially mitigated because email, CRM, and loyalty communications run equally in both treatment and control geographies — they are constants that cancel in the difference estimate. Pros: Captures true omnichannel lift including in-store and BOPIS; immune to privacy signal loss; not dependent on platform infrastructure; produces iROAS estimates for budget allocation.

Cons: Requires substantial spend relative to baseline; minimum 4–8 weeks runtime; geographic media suppression adds operational overhead; spillover effects can bias results in dense store clusters.

3.2 Global Customer Holdout with Cross-Channel Governance

A centrally managed segment (typically 5–15% of the CRM population) is withheld from all owned and paid channels simultaneously under a single governance authority. This is the definitive solution to the overlap problem — it resolves cross-channel contamination by design. The holdout must be stratified by recency, spend decile, category affinity, and geography before any targeting logic is applied; a post-targeting holdout is a biased control group. Hard refresh every 12 months; soft audit every 6 months. New customer cohorts should be allocated to holdout status on arrival to maintain lifecycle representativeness.

Pros: Only methodology that directly resolves the overlap problem; measures cumulative total marketing incrementality; enables nested channel experiments within the exposed population.

Cons: Requires executive sponsorship and cross-channel governance infrastructure; holdout represents suppressed revenue (politically sensitive); does not decompose lift by individual channel without additional within-exposed experimentation.

3.3 Difference-in-Differences & Synthetic Control

These quasi-experimental methods construct a causal counterfactual from observational data without requiring geographic suppression. DiD removes time-invariant group differences and common trends, isolating the media treatment effect. Propensity score-matched DiD — matching unexposed customers on purchase recency, spend tier, and category affinity — can achieve synthetic control quality. Synthetic Control (Abadie et al., 2010 [9]) constructs a weighted composite of unexposed customer segments that mirrors the treated group's pre-period trajectory. Most valuable when a national campaign launched without a pre-planned holdout, or when always-on channels cannot be geographically suppressed. Concurrent channel exposures can be included as DiD regression covariates, partially controlling for the overlap problem.

Pros: Applicable retrospectively; no media suppression infrastructure required; operates at customer level for identified shoppers; produces a visualizable counterfactual trajectory.

Cons: Validity rests on the parallel trends assumption, which is untestable post-treatment; requires rich CRM covariates and analytical expertise; cannot be applied to prospecting campaigns targeting new-to-brand customers with no pre-period CRM history.

3.4 Randomized Platform A/B with Cross-Channel Exclusion Coordination

Extends platform-native lift tools (Meta Conversion Lift, Google Conversion Lift) by exporting the platform holdout list, matching it against the retailer's CRM, and propagating suppression audiences across all concurrent channels before launch. This directly addresses the overlap problem proportional to match rate fidelity. In-store lift can be added via first-party data matching for retailers with high POS loyalty penetration.

Pros: Statistically randomized within the platform; faster to execute than geo testing; enables audience-level lift segmentation (new vs. returning, Pro vs. consumer).

Cons: Match rate limitations reduce exclusion precision; operationally complex — requires coordination across platform, CRM, email, and agency teams; platform methodologies differ across vendors, making cross-platform comparisons unreliable.

4. Tier 2: Speed-Oriented Directional Methods

Tier 2 methods are appropriate when campaign timelines, limited holdout budget, or always-on channel structures make Tier 1 designs infeasible. They trade causal certainty for speed. Interpreted honestly — as

directional signals with stated limitations — they are not inferior instruments but instruments calibrated for different organizational conditions.

4.1 CUPED-Adjusted Pre/Post Analysis

Pre/post analysis compares a metric before and after a campaign, attributing the difference to media. CUPED (Controlled-pre-Experiment Using Pre-Experiment Data; Deng et al., 2013 [11]) substantially improves causal defensibility by regressing post-period outcomes on pre-period behavior: $Y_{adj} = Y - \theta(X - \text{mean}(X))$. In retail datasets, θ typically falls 0.7–0.9, meaning 70–90% of post-period spending is predicted by pre-existing customer behavior. CUPED removes this component, reducing required test duration by 30–50%. Because it operates on aggregate metrics, it naturally captures omnichannel outcomes including in-store revenue. However, without a control group, it cannot distinguish media-driven lift from seasonality or concurrent marketing — results should be interpreted as an upper bound on incrementality.

4.2 Switchback Testing

The unit of randomization is a time window, not a user or geography. Treatment (media on at planned weight) and control (media paused) conditions alternate across defined intervals — typically daily or weekly. The causal estimate is the average difference between treatment-on and treatment-off periods, adjusted for day-of-week and seasonality. Concurrent channel activity is constant across both period types and cancels in the difference estimate, sidestepping the overlap problem. Ideal for always-on channels like paid search where audience holdouts are technically impractical. Carryover effects — where media in treatment periods influences behavior in subsequent control periods — must be modeled or buffered with washout windows.

4.3 Standard Platform Lift Studies

Platform-native tools (Meta Conversion Lift, Google Conversion Lift, Amazon Brand Lift) automatically withhold a random audience percentage, track conversion outcomes, and report the difference as incremental lift. Setup is minimal and results arrive in 2–4 weeks for sufficient-scale campaigns. Critically, these studies measure only digitally observable outcomes by default, systematically understating true omnichannel lift for large-format retailers. Offline conversion uploads extend measurement to in-store purchases at 60–75% match rates. The holdout remains exposed to all other concurrent marketing channels, compressing measured lift and answering "does this platform add value on top of our other channels?" — not "what is this platform's true incremental contribution to revenue?"

4.4 Short-Window Marketing Mix Modeling

Short-window MMM applies regression-based decomposition to 8–16 weeks of data, using Bayesian priors from historical MMM coefficients or industry benchmarks to regularize channel coefficient estimates. It is the only Tier 2 method that natively spans all revenue channels simultaneously — online, in-store, BOPIS — because it operates on aggregate financials, not platform-attributed conversions. It is structurally immune to privacy signal loss and the offline conversion gap. Directional channel contribution rankings are typically stable even when precise point estimates carry wide uncertainty intervals. Results are most reliable when channel spend patterns have sufficient independent variation across the measurement window; highly correlated spend patterns produce multicollinear coefficients.

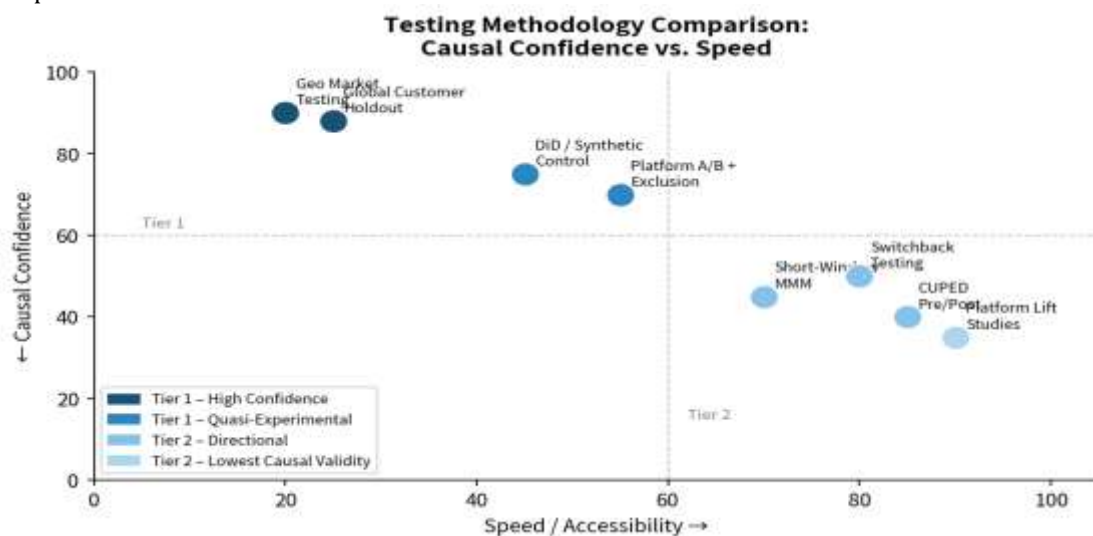


Figure 4. Paid Media Testing Methods: Causal Confidence vs. Speed/Accessibility

5. Decision Framework for Practitioners

The flowchart below guides practitioners through methodology selection based on the measurement question, organizational constraints, and causal confidence requirements. Tier 1 methods apply when causal validity is

worth the investment; Tier 2 methods apply when directional signals are sufficient within planning-cycle constraints.

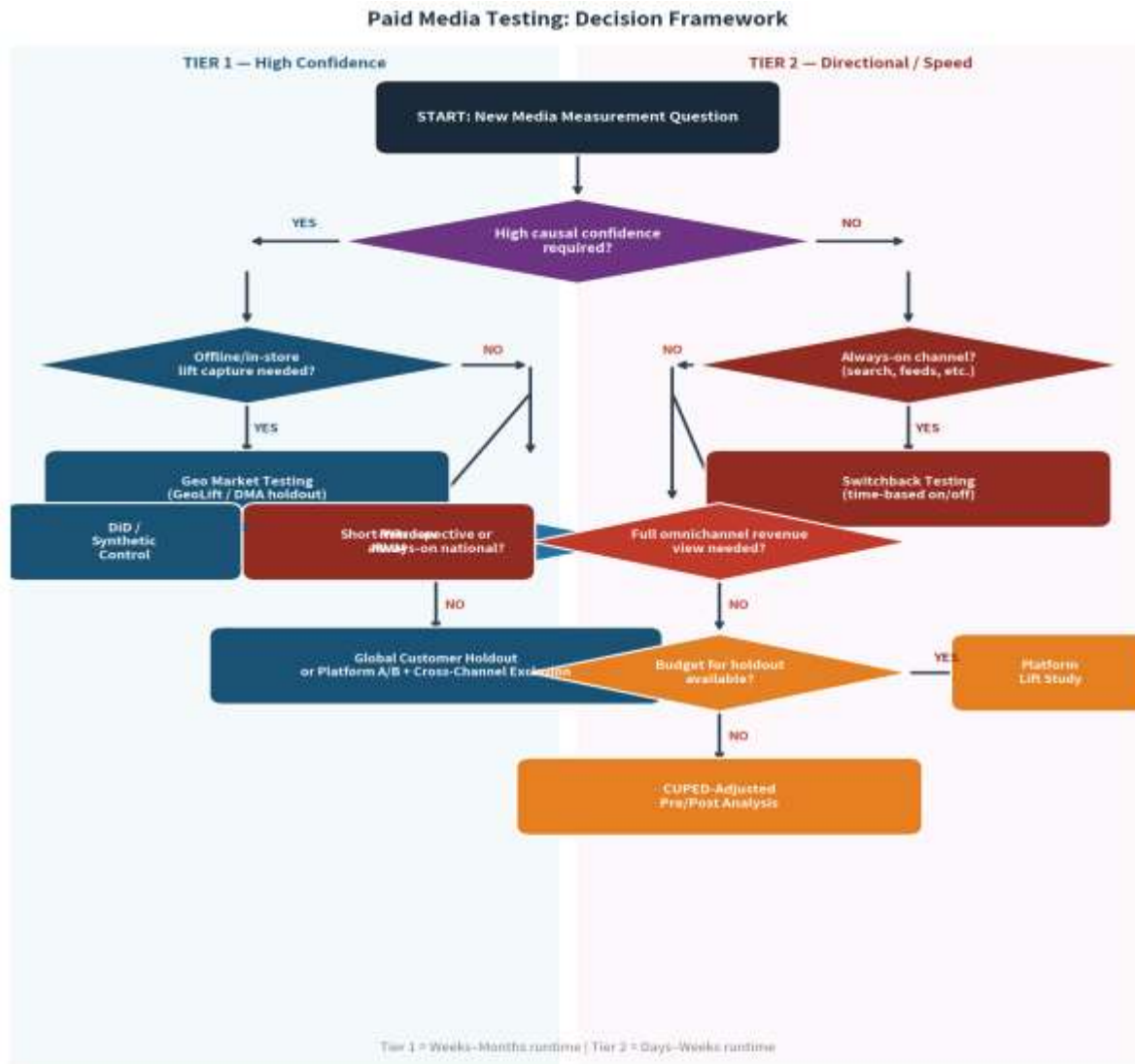


Figure 5. Paid Media Testing Decision Framework — Methodology Selection Flowchart

Methodology Selection Summary

Method	Tier	Offline Lift?	Overlap Handled?	Typical Runtime
Geo Market Testing	1	✓ Yes	✓ Yes (by design)	4–8 weeks
Global Customer Holdout	1	✓ Yes	✓ Yes (definitively)	3–12 months
DiD / Synthetic Control	1	✓ Yes	Partial (statistical)	2–6 weeks
Platform A/B + Exclusion	1	Partial	✓ Yes (coordinated)	2–4 weeks
CUPED Pre/Post	2	✓ Yes	✗ No	2–4 weeks
Switchback Testing	2	✓ Yes	✓ Yes (by design)	2–6 weeks
Standard Platform Lift	2	Partial	✗ No	2–4 weeks
Short-Window MMM	2	✓ Yes	Partial (joint model)	2–4 weeks

6. Conclusion

Large omnichannel retailers operate where measurement errors compound into sustained competitive disadvantage. Yet their structural conditions — offline conversion dominance, walled garden attribution,

privacy signal erosion, and enterprise-scale customer overlap — make rigorous causal measurement genuinely difficult in ways neither academic frameworks nor vendor platforms have fully resolved.

The appropriate response is neither defaulting to the most operationally convenient tool nor pursuing methodological perfection at the expense of agility. It is a deliberate, informed choice. Three organizing principles apply across both tiers: First, the customer overlap problem is not niche — it is a structural feature of enterprise retail that contaminates virtually every channel-specific experiment conducted without centralized holdout governance. Second, offline conversion capture is not optional; measuring only digital outcomes systematically undervalues upper-funnel channels and brand investment. Third, no single methodology resolves all four measurement challenges simultaneously — a portfolio of methods is the mature position.

The paid media measurement discipline is at an inflection point. As user-level tracking degrades under privacy regulation, geo-level and aggregate methods — GeoLift, synthetic control, MMM, switchback testing — are not legacy approaches but the durable infrastructure of modern measurement. Organizations investing now in the governance, data infrastructure, and analytical capability to deploy these methods will maintain measurement continuity as the ecosystem continues to evolve.

References

- [1] eMarketer. (2025). Retail Media Ad Spending Forecast H2 2025. Insider Intelligence.
- [2] Dasika, V. S. (2026). The attribution gap. *International Journal of Leading Research Publication*, 7(5).
- [3] T-Mobile Advertising Solutions / AdExchanger. (2025). The retail media reckoning.
- [4] Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- [5] Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy Online Controlled Experiments*. Cambridge University Press.
- [6] LiftLab / INFORMS. (2026). Omnichannel measurement: Why most ad impact goes unmeasured. LiftLab Research.
- [7] Cutler, M. (2022). Quantifying the impact of Apple's App Tracking Transparency. AppsFlyer.
- [8] Brodersen et al. (2015). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, 9(1).
- [9] Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods. *JASA*, 105(490).
- [10] Angrist, J. D., & Pischke, J. S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.
- [11] Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). CUPED. *Proceedings of WSDM*, 123–132.
- [12] Bojinov, I., & Shephard, N. (2019). Time series experiments and causal estimands. *JASA*, 114(528).
- [13] Chan, D., & Perry, M. (2017). Challenges and opportunities in media mix modeling. Google Research.
- [14] Marchetti, R., & Sepúlveda, T. (2022). GeoLift: Market-level incrementality measurement. Meta Open Source.