GROWTH THAT COMPOUNDS

The incrementality illusion

Why attribution dashboards can report phantom returns while holdout tests find little or no lift, and how to size causal measurement.

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher

PUBLISHED	UPDATED	READING TIME	WORDS
16 August 2026	27 August 2026	15 min	3,274

https://isoglu.com/insights/journal/the-incrementality-illusion/

CONTENTS

MANAGEMENT SUMMARY

Marketing dashboards can report high returns even when a holdout finds little or no incremental sales in some segments. This essay reads three econometric field studies together: Gordon et al. (2019) shows how observational estimates can exceed randomized benchmarks; Blake et al. (2015) separates marginal from inframarginal search demand; Lewis and Rao (2015) shows how variance and expected lift determine the precision a test can deliver. The four-tier framework is a decision aid, not a universal audience cutoff, savings rate, or sample-size rule.

Keywords: Marketing Incrementality · Econometric Attribution · Field Experiments · Ad Spend Efficiency · Statistical Power

The Phantom Return Dilemma

Imagine a commercial dashboard reporting a 420% Return on Ad Spend (ROAS), with Multi-Touch Attribution (MTA) software awarding glowing efficiency scores to retargeting and branded search, while a 25% quarterly marketing budget reduction leaves top-line revenue unchanged. That is the quiet accounting crisis this essay examines.

The numbers reported by the measurement stack were not fraudulent in the narrow sense of forged database rows; they were mathematically accurate summaries of a metric that did not measure what leadership assumed it measured. Much like how every growth budget is a gross number that fails to distinguish asset creation from maintenance, attribution dashboards fail to separate causal impact from natural baseline momentum.

The core pathology of modern commercial analytics is the systematic conflation of intent capture with demand creation. When a prospective customer has already decided to renew a subscription, book a flight, or finalize a B2B procurement contract, modern ad delivery algorithms intercept that buyer at the five-yard line. By displaying an impression or a sponsored search link seconds before the transaction completes, the platform registers a high-confidence conversion. The attribution model marks the dollar as highly productive. But in economic reality, the ad was inframarginal: it spent money to claim credit for a conversion that was already guaranteed to occur.

To understand why commercial teams can pay for phantom marketing returns, and what an evidence threshold requires to establish true causal incrementality, we need three econometric investigations that test observational measurement against experiments:

1. The Observational Breakdown: How advanced non-experimental statistical models fail to recover true causal effects even when given massive demographic and behavioral covariates (Gordon et al., 2019).
2. The Inframarginal Trap: What happens when an enterprise shuts off branded and non-branded search advertising across an entire national market (Blake et al., 2015).
3. The Power Boundary: The mathematical reality of baseline sales volatility that makes true randomized holdout testing unfeasible for mid-market budgets (Lewis & Rao, 2015).

---

1. Why do observational models fail to isolate causal marketing incrementality?

Adtech vendors have presented multi-touch attribution and propensity-score matching as an answer to last-click bias. The premise is intuitive: if an analyst observes rich user-level covariates: browsing velocity, past purchase frequency, geographic tier, device category, demographic markers: statistical matching can construct a valid synthetic control group. By comparing exposed users to “statistically identical” unexposed users, the regression would isolate the true treatment effect.

That premise was tested by Brett Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky in their 2019 Marketing Science study, A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook.

Gordon and his co-authors evaluated 15 large-scale U.S. advertising field experiments comprising 500 million user-experiment observations and 1.6 billion ad impressions. In each experiment, users were randomly assigned to a treatment group (eligible to see the advertiser’s campaign) or a control group (shown public-service an-

nouncements or unexposed). Because assignment was randomized at massive scale, the difference in conversion rates between treatment and control represented the true, unconfounded causal effect of the advertising.

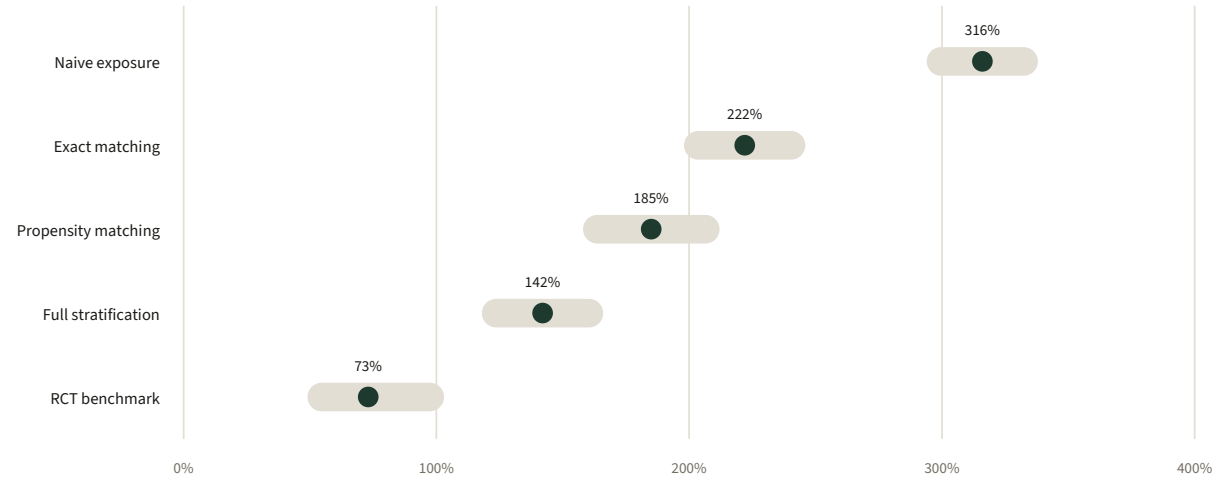
The researchers then applied the industry's full arsenal of non-experimental methods: including ordinary least squares (OLS) regression, exact matching, propensity score matching (PSM), and machine-learning-based stratification: to the exact same datasets, conditioning on Facebook's vast internal user profiles and browsing histories.

The Scale of the Error

The results expose a large gap between observational estimates and the randomized benchmark:

- Magnitude and sign errors: In 7 of the 14 campaigns with checkout-conversion outcomes, observational point estimates were off by more than a factor of three; in multiple cases, they also had the wrong mathematical sign. The paper also reports cases where an observational model found positive, statistically significant returns while the randomized trial found no effect.
- No universal correction: These are study-specific comparisons, not evidence for applying a fixed 3x or 4x correction to every campaign.
- Failure of Rich Covariates: Even when models conditioned on deep behavioral variables (including historical platform activity, device engagement, and prior conversion signals), the observational bias was not eliminated.

Figure 1 Estimated treatment effect (ATT) on checkout conversions across model specifications



Gordon et al. (2019), Marketing Science 38(2), §7.

The underlying mechanism is activity bias (Lewis et al., 2011). Consumers who are actively in-market for a product exhibit high digital activity: they visit review sites, interact with related content, and spend time online. Ad auction algorithms are engineered to identify this exact behavioral surge and place bids accordingly.

Consequently, high-intent consumers receive far more ad impressions simply because they are about to buy. Observational models observe a high correlation between ad exposure and transactions, and erroneously assign credit to the ad. In reality, ad exposure and purchase are joint downstream consequences of an unobserved latent variable: pre-existing consumer purchase intent.

Regression adjustment alone cannot identify that causal effect when intent remains unobserved. An exogenous design, such as withholding the ad through randomized assignment, is one way to create the needed contrast.

---

2. What did the eBay search blackout prove about inframarginal ad spend?

While Gordon et al. analyzed display and social advertising, an even starker demonstration of intent-harvesting occurs in paid search engine marketing (SEM).

In their 2015 Econometrica paper, Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment, Thomas Blake, Chris Nosko, and Steven Tadelis examined the causal returns of search advertising at eBay. At the time, eBay was one of the largest digital advertisers on earth, spending hundreds of millions of dollars annually bidding on brand keywords (e.g., “eBay”, “eBay shoes”) and non-brand keywords (e.g., “used Gibson guitar”, “memory card”).

Standard attribution reporting at eBay indicated that search ads were an indispensable growth engine, returning hundreds of percent in net ROI. To test whether these returns were real, the economics team executed two massive experimental interventions.

Experiment 1: The Branded Keyword Blackout

In the first test, eBay systematically suspended all paid search advertising on branded terms across major search engines for entire geographic regions (DMAs) in the United States while leaving organic listings intact.

The result was near-total substitution:

> **Organic Recapture Rate ~99.5%**

The moment the sponsored link at the top of the search results page disappeared, users simply clicked the unpaid organic link located immediately below it. Total click traffic to eBay remained virtually unchanged, and purchases remained statistically flat.

The branded search campaign had been spending millions of dollars per month purchasing navigation clicks from users who had already typed the company’s name into their browser. In this test, the ad was extractive: it transferred consumer surplus to the search engine without producing a measurable incremental purchase.

Experiment 2: The Non-Brand Search Experiment

Proponents of paid search argued that while branded search might be inframarginal, generic non-brand keywords represented true customer acquisition. Blake, Nosko, and Tadelis tested this by halting all non-brand search ads across 65 randomly selected Designated Market Areas (DMAs) for a period of several months, comparing sales trajectories against control markets.

The findings uncovered a critical structural heterogeneity:

1. Zero Lift for Frequent Buyers: For users who had purchased from eBay at least once in the prior year, paid search ads had no statistically measurable effect on purchasing behavior. These customers were already aware of the platform and navigated directly or via organic channels.
2. Positive Lift for Marginal Users: For genuinely new and infrequent buyers (< 1 transaction in prior year), non-brand search ads generated a statistically significant lift in purchases.
3. Negative Channel ROI: Because frequent, high-volume buyers accounted for the overwhelming majority of search queries and ad click costs, the budget consumed by inframarginal buyers completely wiped out the economic gains from acquiring new users. The aggregate Return on Investment for the non-brand search channel was negative.

Table 1 Non-brand search effectiveness by consumer cohort

Consumer segment	Result in this experiment
Frequent users (at least one purchase in the prior year)	No statistically measurable purchase effect
New and infrequent users	Positive and statistically significant purchase lift
Aggregate non-brand search channel	Negative return in the study's aggregate calculation

Blake et al. (2015), Econometrica 83(1).

The strategic lesson for commercial operators is profound: an ad channel can be simultaneously effective for a specific sub-population and value-destroying in the aggregate. When targeting algorithms cannot restrict delivery exclusively to uninformed, marginal buyers, high-intent inframarginal users soak up the budget, transforming a profitable acquisition tool into an expensive subsidy.

3. Why are statistical power boundaries unfavorable for digital lift studies?

Faced with the failure of attribution models and the reality of inframarginal leakage, the instinctive executive reaction is: "We must mandate randomized holdout tests across every channel."

Here, commercial leadership collides with a rigid mathematical barrier documented by Randall Lewis and Justin Rao in their 2015 Quarterly Journal of Economics paper, The Unfavorable Economics of Measuring the Returns to Advertising.

Lewis and Rao demonstrate why measuring the return on advertising is fundamentally harder than measuring outcomes in clinical medicine, industrial engineering, or software conversion optimization.

The Mathematics of Baseline Volatility

In clinical drug trials, the baseline variation in patient health markers is relatively compact, and treatment effects are often substantial. In commercial advertising, the reverse is true:

1. Individual purchasing behavior is extremely volatile (sigma is large): On any given day, an individual customer's probability of making an enterprise purchase or retail checkout is heavily skewed, characterized by massive variance and fat-tailed purchase sizes.
2. The marginal causal lift of an ad campaign is modest (delta is small): A highly successful advertising campaign rarely increases conversion rates by 50%; it increases a baseline conversion rate from, say, 1.00% to 1.08%: an incremental lift of 8 basis points (delta = 0.0008).

At fixed design choices, a useful planning relationship for the required sample size N per arm is:

$$N \text{ scales with } (\sigma / \delta)^2$$

The constant depends on the outcome, allocation, power, significance level, clustering, spillover and analysis plan. Because the noise-to-signal ratio is squared, smaller expected lifts can increase the required sample sharply.

Table 2 Planning the sample size for causal advertising lift

Expected relative lift	Planning implication	Inputs required before calculating
Larger	Required N can be lower, all else equal	Baseline outcome, variance, allocation, power, alpha, clustering and spillover
Moderate	Required N can be substantially larger	The same inputs, plus a decision-relevant minimum effect
Very small	The design may become impractical	A pre-specified precision target, opportunity-cost limit and stopping rule

Author's planning worksheet based on Lewis & Rao (2015), Quarterly Journal of Economics 130(4). No single set of outcome or design inputs is assumed here.

Lewis and Rao examined 25 field experiments executed by major US retail and digital brands. They calculated that the median 95% confidence interval for campaign ROI spanned this range:

$$\text{Median 95\% Confidence Interval for ROI} = [-35\%, +175\%]$$

In plain terms: even after the six-figure expenditure and large user counts reported in this set of experiments, the estimate could be so noisy that leadership could not distinguish between a substantial loss and a large positive return.

The Mid-Market Dilemma

At very high volumes, continuous experimentation may be feasible. For a mid-market enterprise, an e-commerce brand, or a B2B software company with long-cycle outcomes:

- Power may be limited: The available conversion volume may not support a precise estimate of a small lift over baseline noise.
- Holdouts have an opportunity cost: Withholding advertising sacrifices potential margin, so the value of information should be weighed against the cost and duration of the holdout.

Commercial leaders who demand “pure experimental proof” for mid-sized budgets are requesting an instrument that physical mathematics refuses to provide. Understanding why growth compounds requires recognizing that durable baseline expansion stems from cumulative product and brand capital, not micro-optimized attribution dials.

4. The Executive Decision Architecture

If observational attribution can overstate returns, search ads can harvest inframarginal demand, and randomized holdout tests can be difficult at lower volumes, how should a commercial organization govern its growth capital?

The resolution is to abandon the single-dashboard fantasy and deploy a four-tier measurement governance protocol matched to the structural physics of each marketing channel.

Table 3 The four-tier commercial measurement governance framework

Tier & Scale	Applicable Channels	Mandated Measurement Methodology
Tier 1: High scale	Display, social or paid search where volume and geography support testing	Randomized geo-holdouts and ghost ads: Measure incremental cost per acquisition; do not let platform attribution govern alone.
Tier 2: Macro mix	Brand, CTV, podcast and other aggregated awareness channels	Calibrated Bayesian media mix modeling: Use time-series models anchored by periodic experimental evidence where feasible.
Tier 3: Low volume	B2B pipeline, account-based marketing, niche campaigns or long-cycle outcomes	Unit-economic payback bands: Govern contribution margin and payback, and use causal designs when their precision and cost are acceptable.
Tier 4: Defensive	Branded search and bottom-funnel retargeting	Measured defensive restrictions: Use exclusions, bid tests and frequency or recency caps set as experiment parameters.

Author's synthesis of the cited econometric literature.

Tier 1: High-Volume Direct Channels: Randomized Regional Holdouts

At high volume, treat observational attribution as a hypothesis generator rather than the governing metric. There is no universal audience cutoff: the right design depends on baseline conversion, variance, geography, spillover and the cost of withholding spend.

- Geo-Holdout Matched Markets: Rather than splitting individual users, partition geographies into matched test and control pairs based on historical sales velocity, seasonality and likely spillover.

- Ghost Ad Technology: Use synthetic control ad technology (Johnson et al., 2017) to observe when an ad would have been delivered to a control user without actually serving a competing PSA, shrinking the variance of the counterfactual estimate.
- Governing Metric: Optimize exclusively to incremental Cost Per Acquisition (iCPA):

$$iCPA = (Ad\ Spend_{Test} - Ad\ Spend_{Control}) / (Conversions_{Test} - Conversions_{Control})$$

If a geo-holdout iCPA is higher than the gross-profit threshold after uncertainty and spillover are considered, scale the channel down regardless of what the platform dashboard claims. The threshold is a business decision, not a universal benchmark.

Tier 2: Macro Brand & Broadcast: Calibrated Media Mix Modeling

For multi-channel brand investments where individual-level tracking is impossible, commercial organizations should deploy modern open-source Bayesian Media Mix Modeling (e.g., lightweight state-space models with ad-stock and saturation transformations).

Crucially, raw regression MMM is vulnerable to the same endogeneity traps as attribution (firms spend more during holiday peaks, causing models to assign organic holiday demand to ad spend). To solve this:

1. Experimental Calibration: Treat periodic geo-holdout tests as prior distributions within the Bayesian model.
2. Anchor Priors: If an RCT estimates a lower incremental lift coefficient than a regression model, use that experimental estimate to inform the prior in the broader macro mix model. The coefficients are illustrative, not portable values.
3. Decompose Base vs. Lift: Explicitly model the un-promoted organic baseline to prevent macro economic drift from being misattributed to media spend.

Tier 3: Low-Volume & B2B: Unit-Economic Bounding Rules

When user volumes make a well-powered test difficult, causal precision may require a different design or a longer observation window. For B2B and niche enterprise teams:

- Stop treating attribution as allocation evidence: If a multi-touch system assigns percentage weights across a complex enterprise journey, use the output as a descriptive signal unless it has been calibrated against an experiment. Do not confuse a dashboard subscription with causal proof.
- Set cash payback windows explicitly: Establish unit-economic guardrails tied to gross margin, cash constraints and the sales cycle. The right duration is a management choice for the business, not a universal twelve-month or four-month benchmark.
- Blended CAC Governance: Evaluate marketing spend as a collective portfolio against net new annual recurring revenue (ARR) additions, treating individual channel touches as qualitative engagement indicators rather than causal proof.

Tier 4: Branded Search & Retargeting: Defensive Rules

Branded search and aggressive site retargeting are useful places to test for inframarginal waste. Commercial leadership should impose structural constraints that are measured, not copied as universal rules:

1. Negative Audience Matching: Mechanically exclude all existing customers, active subscribers, and employees from paid search and retargeting pools via hashed first-party CRM lists.
2. Branded Search Extraction Tests: Periodically lower brand keyword bids to zero in test regions. If organic recapture is high and conversions do not fall beyond the study's uncertainty range, reduce or eliminate the defensive spend.
3. Retargeting frequency tests: Set frequency and recency caps as experiment parameters, then review incremental conversion and margin. The right cap depends on the audience, buying cycle and creative; two impressions per week and seven days are not universal laws.

What is the economic payoff of causal measurement integrity?

The persistence of the incrementality illusion is not a technical mystery; it is an organizational alignment problem.

Ad platforms have every structural incentive to report total touched conversions rather than incremental lift. Ad agencies whose compensation is pegged to a percentage of media spend have no incentive to recommend shutting off high-ROAS branded search campaigns. Marketing executives presenting quarterly decks to the board face immense pressure to display clean upward-trending efficiency charts.

Moving from observational attribution to causal incrementality is painful because it can materially lower reported marketing efficiency. The size of that gap is channel-, audience- and design-specific, so a single 50% to 75% reduction is not a general result.

The transition can release budget from transactions that were already going to occur, but the amount must be measured in the organisation's own tests. Redirecting that released budget toward genuinely incremental acquisition is a decision to validate, not a guaranteed highest-return reallocation.

Causal measurement does not promise that every marketing dollar will look brilliant on a dashboard. It promises something far more valuable: that when your enterprise writes a check for growth, it is paying for a customer it did not already own.

Where are the experimental boundaries of incrementality testing?

Boundary. The article sets decision limits for observational attribution; it does not estimate a universal savings rate, audience cutoff, or exposure frequency. Calibrate withholding, spillover, margin, and test capacity against the channel being evaluated.

Evidence base. The analytical frame also draws on these additional sources: Bronnenberg et al. 2012. The links identify the exact works; they support the mechanisms and boundary conditions discussed here, not every claim in isolation.

Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. Econometrica, 83(1), 155–174. https://doi.org/10.3982/ECTA12423

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science, 38(2), 193–225. https://doi.org/10.1287/mksc.2018.1135

Johnson, G. A., Lewis, R. A., & Nubbemeyer, E. I. (2017). Ghost ads: Improving the economics of measuring online ad effectiveness. Journal of Marketing Research, 54(6), 867–885. https://doi.org/10.1509/jmr.15.0297

Lewis, R. A., & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. The Quarterly Journal of Economics, 130(4), 1941–1973. https://doi.org/10.1093/qje/qjv023

Lewis, R. A., Rao, J. M., & Reiley, D. H. (2011). Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. Proceedings of the 20th International Conference on World Wide Web, 157–166. https://doi.org/10.1145/1963405.1963431

Bronnenberg, B. J., Dubé, J. P., & Gentzkow, M. (2012). The evolution of brand preferences: Evidence from consumer migration. American Economic Review, 102(6), 2472–2508. https://doi.org/10.1257/aer.102.6.2472

END OF ESSAY

HOW TO CITE THIS ESSAY

Isoglu, S. (2026, August 16). The incrementality illusion. Sinan Isoglu. https://isoglu.com/insights/journal/the-incrementality-illusion/

KEEP READING

Growth that compounds

What is marketing efficiency ratio (MER)? A blended signal, not a causal answer

https://isoglu.com/insights/journal/what-is-marketing-efficiency-ratio/

Growth that compounds

What is revenue growth management? Price, volume, mix, and margin in one system

https://isoglu.com/insights/journal/what-is-revenue-growth-management/

Growth that compounds

Long customer relationships can be low-profit

https://isoglu.com/insights/journal/long-customer-relationships-can-be-low-profit/

Sinan Isoglu, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher

Online version: https://isoglu.com/insights/journal/the-incrementality-illusion/