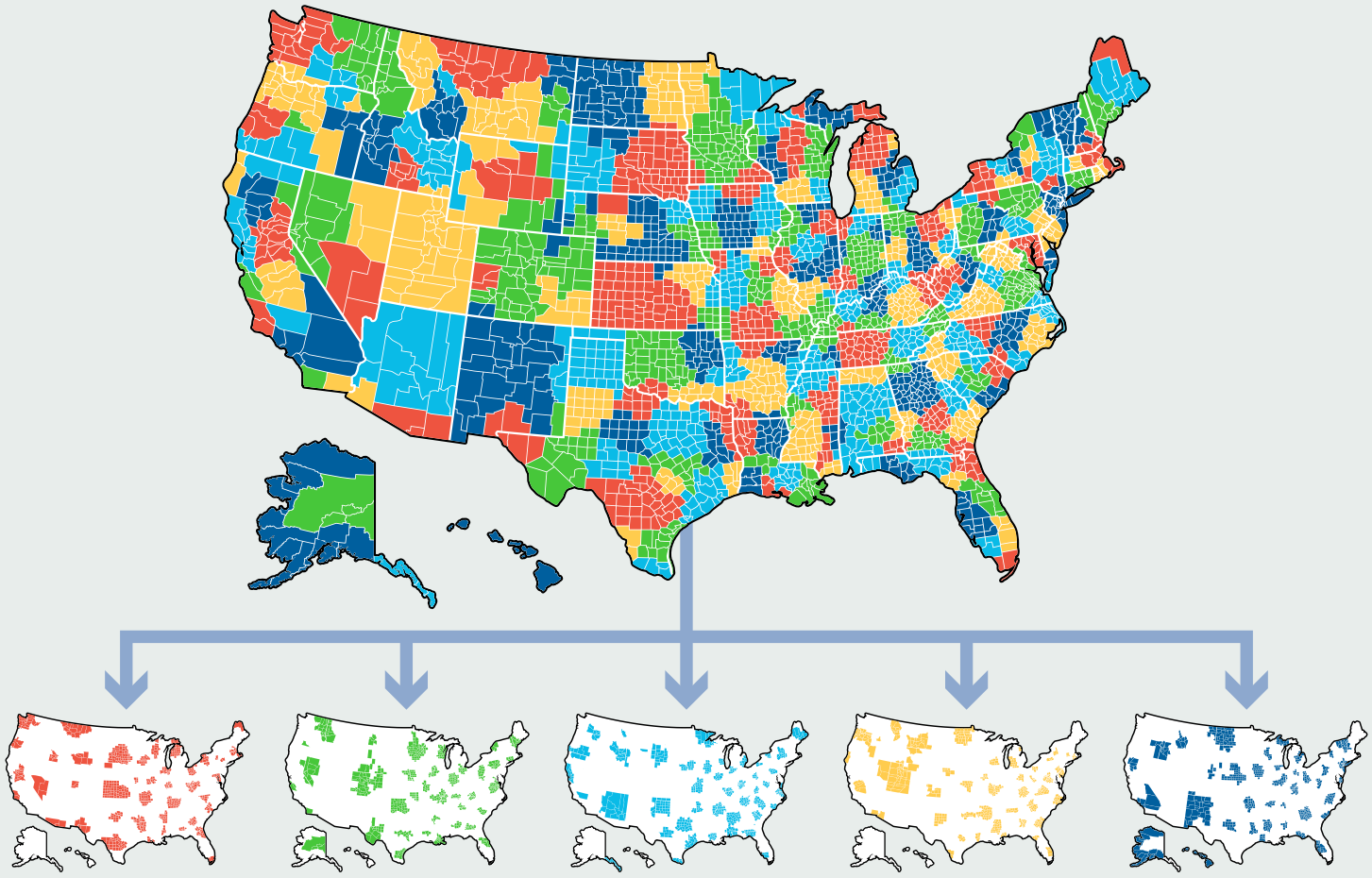
CENTRALCONTROL

TURN AD PERFORMANCE CERTAINTY INTO MARKET ADVANTAGE

HOW TO DESIGN A GEOGRAPHIC RANDOMIZED CONTROLLED TRIAL

A DETAILED GUIDE TO "ROLLING THUNDER" AND OTHER TRANSPARENT, UNBIASED AND REPLICABLE TECHNIQUES FOR MEASURING INCREMENTAL ADVERTISING IMPACT WITH EXPERIMENTAL DESIGN



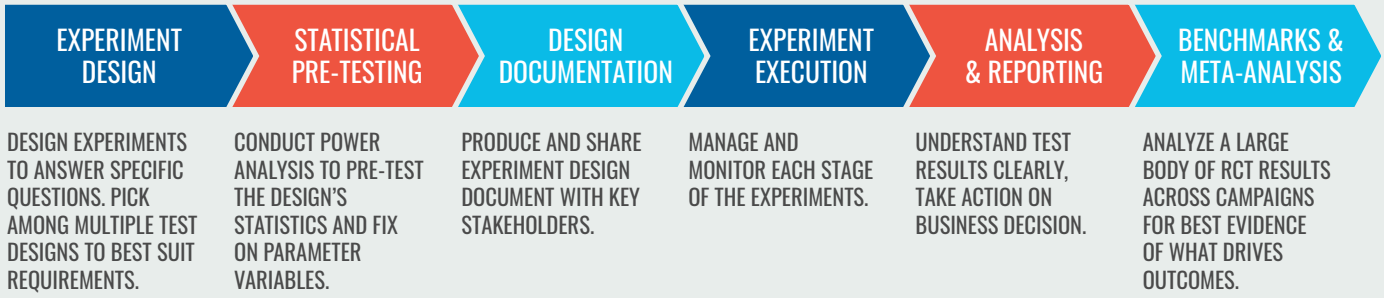
CONTENTS

INTRODUCTION

At Central Control, we have written several articles[1] advocating for advertisers to measure incremental return on ad spend (iROAS) by using large-scale geographic randomized controlled trials (RCT). This article provides a detailed explanation of how to design and execute such experiments.

Those previous writings covered extensively why geo experiments offer the best evidence of what's working in your marketing mix compared to user-level experiments, quasi-experiments (e.g., synthetic control, matched market tests), attribution, and other observational methods. This piece focuses on how: the practical implementation of high-quality large-scale geographic RCTs.

EXPERIMENT PROCESS WORKFLOW



▲ This workflow summarizes the end-to-end process for running advertising experiments. The final stage – benchmarking – compiles results from all experiments into a growing evidence base. As advertisers, agencies, and media companies accumulate dozens, then hundreds, or even thousands of tests, these benchmarks become a powerful source of insight. They reveal what truly works in advertising and provide expected effect sizes by media channel, ad format, partner, product category, and other key dimensions.

[1] Advertisers seeking accurate ROAS should use large-scale, randomized geo tests

Gain market share bydisrupting bad ad measurement

Enterprise AI is coming, and it's about to learn all the wrong lessons about marketing effectiveness

FIRST STEPS

Before designing the experiment, several foundational questions should be addressed.

DETERMINE THE RESEARCH QUESTION

STATE YOUR BUSINESS OBJECTIVE AND THE DECISION TO BE SUPPORTED

Experiments are not as difficult or expensive as many marketers perceive or portray them, but they do require resources and calendar time, so run them only when the answer will drive a meaningful business decision. Geo experiments are almost always “intent to treat”(ITT) designs, where treatment is tracked at the geographic level rather than the individual level. You know which geographic populations were intended to receive treatment, not which individuals actually saw the ads. This means your media investment must be substantial enough to detectably move KPIs at the geographic level. This technique isn’t suited for relatively small campaigns or very rare conversions.

Start by measuring sales effects for your largest media channels to validate and calibrate coefficient assumptions in your marketing mix model. From there, having laid a baseline of channel-level validation and gotten a grounding in best practices for conducting experiments, researchers can branch out into nuances of the mix, such as the effectiveness of particular media partners, subchannel tactics, media formats and so forth.

Most importantly, identify the specific business decision that will be made based on the results, with marketing, analytics and finance working in concert. A well-constructed geo experiment that doesn’t inform action is a waste of resources.

Example business objectives:

- Validate that the impact of a marketing channel, such as Search (or Television, Social, Retail Media, etc.) matches our marketing mix model’s (MMM) coefficients, and if not, investigate (and experiment) further with a view to calibrate the MMM accordingly and reevaluate investment levels
- Measure iROAS of a media partner, such as Meta (or Google, Spotify, Disney, etc.) as a channel partner and renegotiate or reallocate if results are unfavorable
- Evaluate iROAS of a tactic, such as Retargeting (or Lookalike Targeting, Branded Search Keywords, Geo-fencing, etc.) and adjust investment accordingly

DEFINE KEY EXPERIMENT VARIABLES

Next, we define the key parameters that will structure the experiment.

INDEPENDENT VARIABLE: THE MEDIA INVESTMENT

In an experiment, the independent variable is what you modify between test and control groups to measure its effect on the outcome variable. For advertising experiments, this means the ad campaign or media channel being tested, whether branded search, social media, television, or another tactic.

Since we're designing geographic experiments, it's crucial to verify that your chosen medium can reliably target different geographic units sufficiently for your testing objectives. This capability varies significantly across media types and platforms. Examples, as of this writing: Linear TV can be tested effectively using DMA targeting through providers like Ampersand or Nexstar, but individual networks like CBS or AMC may lack sufficient scale or targeting granularity. Spotify enables geo-targeting for streaming music and run-of-network podcasts but not for host-read spots on specific shows. Amazon supports geo-targeting for display and video on-site and their DSP but not for sponsored product keywords that comprise most retail media budgets. ZIP codes are not a stable testing unit for most media.

DEPENDENT VARIABLE: KEY PERFORMANCE INDICATOR (KPI)

The dependent variable is the outcome you're measuring, ideally sales. The fundamental assumption is that this outcome is influenced by the independent variable: running advertising increases sales. The experiment aims to quantify this causal relationship by comparing against the counterfactual baseline sales rate in the control group absent the advertising stimulus.

Sales is the preferred KPI as it directly measures iROAS, customer acquisition cost (CAC), and cost per acquisition (CPA), all metrics that should reflect true incrementality. Other outcomes such as foot traffic, search activity, web visitation, and brand lift can also be measured, but should be measured with RCT to know if the ads are having a true causal effect.

Critically, ensure you have reliable KPI data that aligns with the experimental structure. For geographic experiments, postal codes in customer databases are ideal – no device graphs, tracking pixels, or clean rooms required. Whether from first-party CRM systems or partner panels, ZIP codes easily roll up to most standard geographic units, allowing straightforward determination of whether each customer belonged to test or control groups at the time of conversion.

EXPERIMENT UNIT: GEOGRAPHIC REGIONS

That doesn't work with custom units such as radial or hexagonal zones. In RCTs, "experiment units" are the smallest entities to which treatment is independently applied and which are subject to randomization. For advertising, these could be individuals, devices, households, or geographic regions.

As explained in our article "Advertisers seeking accurate ROAS should use large-scale, randomized geo tests,"[2] user- and household-level targeting provides false precision. User-level tracking accuracy has generally been overstated by vendors, and Apple's App Tracking Transparency (ATT) and Safari browser have further degraded online tracking capabilities.

[2] Advertisers seeking accurate ROAS should use large-scale, randomized geo tests

The internet's promise of 1:1 targeting was never fully realized. Cookies are inherently unstable, regularly deleted and lacking persistent identity. Industry responded with complex identity graphs and clean rooms. Users pushed back with ad blockers, while governments enacted privacy laws like GDPR. As a result, when marketers attempt to match campaign-exposed audiences to first-party outcome data, match rates typically range from 50-60% or less.

Even at 80%, this introduces too much noise to reliably detect typical advertising effects of 1-5% with statistical confidence.

This explains the rise of geographic experiments. Using geo regions as experimental units forms what's known as cluster-randomized trials (CRT), where geographic clusters of individuals are randomly assigned to treatment conditions.

Geographic experiments offer multiple advantages. They sidestep privacy concerns by eliminating the need for personally identifiable information. Sales and other KPIs align naturally with this structure through ZIP-coded transaction data from CRM systems, without privacy implications or technical overhead.

Most media can accurately target large geographic regions, particularly Nielsen's designated market areas (DMAs). While postal codes, cable zones, core based statistical areas (CBSAs), and states are alternatives, each presents challenges. Postal codes would be ideal, but as we've written in AdExchanger[3], media companies currently infer them from IP addresses with high error rates or accumulate multiple ZIPs per user account, making them unreliable for experimentation without additional system enhancements

Geographic regions suitable for experimentation must be mutually exclusive and sufficiently isolated to minimize spillover between test and control groups[4], upholding the experimentation best practice known as Stable Unit Treatment Value Assumption (SUTVA). Ideally, units should also be collectively exhaustive (MECE: mutually exclusive and collectively exhaustive).

DMAs meet these criteria well. They're mutually exclusive and collectively exhaustive across the US. Most people don't commute between DMAs – while NYC and Philadelphia DMAs border each other, their populations generally live and work within one or the other. With 210 US DMAs, randomization can balance the many factors affecting regional sales: income levels, education, competitive mixes, weather, and more.

Critically, virtually all media types can be targeted reliably by DMA in the US, a practice established since the 1950s.

Outside the US, metro areas or postal-code clusters typically work best.

Bear in mind, we're talking here about the RCT subclass of cluster randomized trials, so don't confuse this approach with matched-market tests (MMT) or synthetic control method (SCM). Those are forms of quasi-experiments, a scientific designation that means "not as good as real experiments," defined by a lack of randomization of assignments of test and control arms. This paper focuses on how to conduct geo RCTs for iROAS measurement and doesn't get into why CRT provides categorically better evidence than MMT or SCM. For more on that discussion, refer to the articles cited in Footnote 1.

For this paper, we will focus on large-scale randomized geographic experiments using DMAs as the experimental unit.

[3] ZIP Codes: The Simple Fix For Advertising ROI Measurement

[4] Measuring Ad Effectiveness Using Geo Experiments" (Vaver, Koehler,

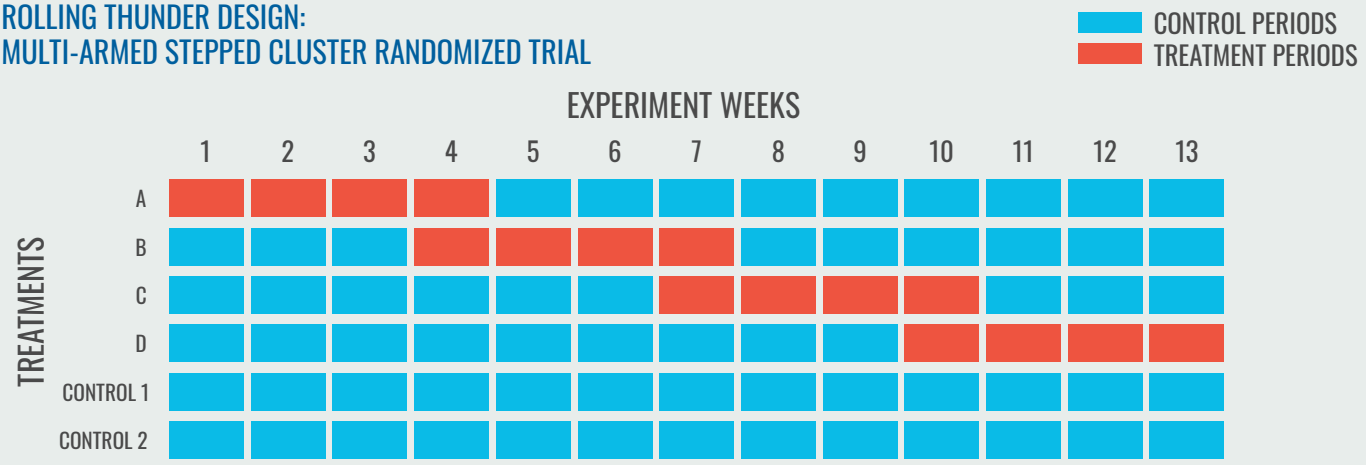
SELECT DESIGN TYPE

There are myriad types of designs for running experiments. We have already narrowed to some key parameters, such as that we are focusing on using geographic regions for the experiment units and also that we are using an intent-to-treat (ITT) design, meaning that we do not need to track who in the test group was actually treated with the campaign exposure, and that treatment is withheld from everyone in the control group. We will briefly discuss here a few other more design details, but settle on one for the rest of this paper.

Different business questions give rise to different experimental design choices. Those include:

- **'Rolling Thunder': Multi-armed, stepped CRT:** This design offers greater statistical power than a traditional two-celled test by generating many independent contrasts across time and treatment arms[5]. Each group enters treatment in a staggered sequence and remains in for a fixed duration, creating repeated observations both within and across DMAs. This structure increases the precision of the causal estimate while reducing sensitivity to outliers or time-based shocks. The design is particularly well-suited for cessation testing: by phasing off the media gradually, advertisers can closely monitor sales impact, minimize the risk of revenue loss, and halt the test early if needed.

ROLLING THUNDER DESIGN:
MULTI-ARMED STEPPED CLUSTER RANDOMIZED TRIAL



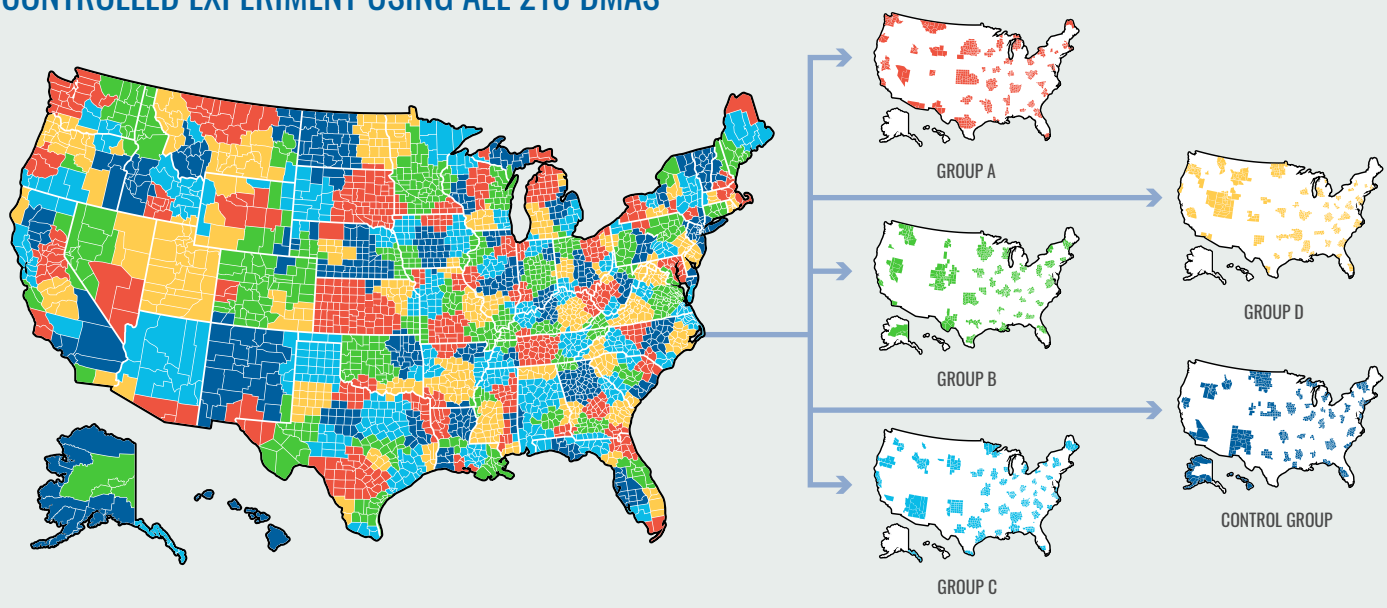
▲ This diagram shows a test structure for evaluating the sales impact of media. It is useful for testing in a new channel but also as a cessation test for pausing an always-on media channel the advertiser believes is essential to sales. We have nicknamed this design type "Rolling Thunder." In this example, four treatment groups are turned off in staggered 4-week windows to minimize overall disruption. At any given time, only a subset of the country is in test: in this example, typically one-sixth, and at most one-third (in weeks 4, 7, and 10). Multiple control groups and spacing between test flights reduce risk. Rules for halting the test based on observed negative impact should be pre-specified during the design phase. The configuration, including number of test arms, treatment duration, and control size, should be guided by pre-test power analysis to ensure detectability of the expected effect size.

[5] Stepped wedge designs could reduce the required sample size in cluster randomized trials" (Woertman, de Hoop et al., 2013)

- **Targeting Test:** To see if the iROAS of a targeting technique pays off, this three-armed design includes one test group for targeting, one test group for minimally targeted media using the same creative, and one unexposed control group.
- **Media Bake-Off Trial:** When comparing two media channels against one another for the superior ROI, a three-armed test, one for medium A, one for medium B and a control group
- **Media Bake-Off Plus Interaction Effect:** A four-armed test compares iROAS of two media properties plus their interaction effect, with one exposed to medium A, one medium B, one both, and one unexposed control group, using factorial design and analysis
- **Intensity Trial:** For determining optimal media weight and calibrating diminishing returns, a multi-armed test where different groups receive varying media levels (e.g., 50%, 100%, 200% of baseline spend) plus an unexposed control. Unlike cessation tests which measure what's lost when media stops, intensity trials identify the point of maximum efficiency by testing increased investment levels. This design reveals whether doubling spend doubles impact or if returns diminish, critical for optimizing budgets rather than just validating them.
- **Parallel Two-celled Test:** This is the most basic design for a lift test, answering the question of how much incremental effect an ad campaign/channel/publisher/tactic or other ad strategy as on a desired outcome, such as sales, with one test group and one control. Statistically weaker than Rolling Thunder design: recommended only for estimated effect size great than 5%

Practitioners may opt to have different ratios of the various arms of the experiments, but our recommendation is to use equal probability in assigning arms whether that's 50/50 for two-armed tests or likewise for any other number of splits. Uneven ratios between test and control weaken the statistical power of the test and risk skewing the sample between the arms.
The analysis for each of the designs mentioned above functions similarly.

EXAMPLE OF MULTI-ARMED GEOGRAPHIC RANDOMIZED CONTROLLED EXPERIMENT USING ALL 210 DMAS



DESIGN THE EXPERIMENT

With the business problem defined and the experimental approach selected we can now proceed to the detailed design.

DEFINE EXPERIMENT PARAMETERS

EFFECT SIZE

Specify the effect size you intend to detect, typically expressed as a percent lift in the primary KPI (e.g., sales). This value should reflect what the business would consider meaningful, based on past campaign benchmarks, modeling outputs (e.g., marketing mix models or multitouch attribution models), or practical judgment. For example, an expected lift of ~3% may be a reasonable planning target for most large consumer brands.

It may feel odd to have to estimate this quantity, since the point of the experiment is to determine its value. Don't think of this as a forecast. Instead, think of this as the minimum lift that would be worth detecting based on your business. A point of departure for this, if lacking any better guidance, is what would be the minimum lift required to generate positive iROAS for the campaign. The lift estimate is an input to power analysis that anchors the design in real-world relevance.

Power analysis will assess whether the proposed design can detect this effect with sufficient probability and may show that the setup is sensitive enough to detect even smaller effects. Avoid designing the experiment to detect only the absolute minimum effect you think might occur, as this can overfit assumptions and raise the risk of an underpowered test.

Where feasible, simulate power across a range of plausible effect sizes to assess robustness and trade-offs in detectable lift. Given that this is a how-to paper concerning experimental design, this material gets at least a bit more technical as it proceeds. We assume the reader has some familiarity with modern statistical techniques, but we trust that non-technical readers should be able to follow the basic concepts as we continue.

CONFIDENCE LEVEL

Specify the confidence level that will be used to test for statistical significance. The standard is 95%, corresponding to a significance threshold (α) of 0.05, although 90% confidence is often acceptable for marketing purposes.

This determines the risk of a false positive ("Type I error"), i.e., concluding there was an effect when there wasn't. A 95% confidence level implies that, under the null hypothesis (the opposite of the hypothesis), there's a 5% chance of incorrectly declaring a statistically significant result.

TEMPORAL WINDOWS

Pre-Period (Baseline) Length

Set a baseline window – eight weeks is usually sufficient, depending on sales cycles – to normalize each DMA's trend before treatment. This period is used for trend stabilization and covariate adjustment. Justify that it is long enough to smooth out short-term noise but recent enough to reflect current conditions. Avoid choosing periods that include holidays or anomalies unrepresentative of the test window.

Exposure Period

Determine the active test duration based on expected consumer response. For high-reach media (e.g., TV or digital display), a 4-6 week exposure is often sufficient. For lower-latency channels (e.g., search, retail media), a shorter window may suffice. Ensure media weight and pacing can be reliably sustained during this period. This can be informed during the simulations of the power analysis.

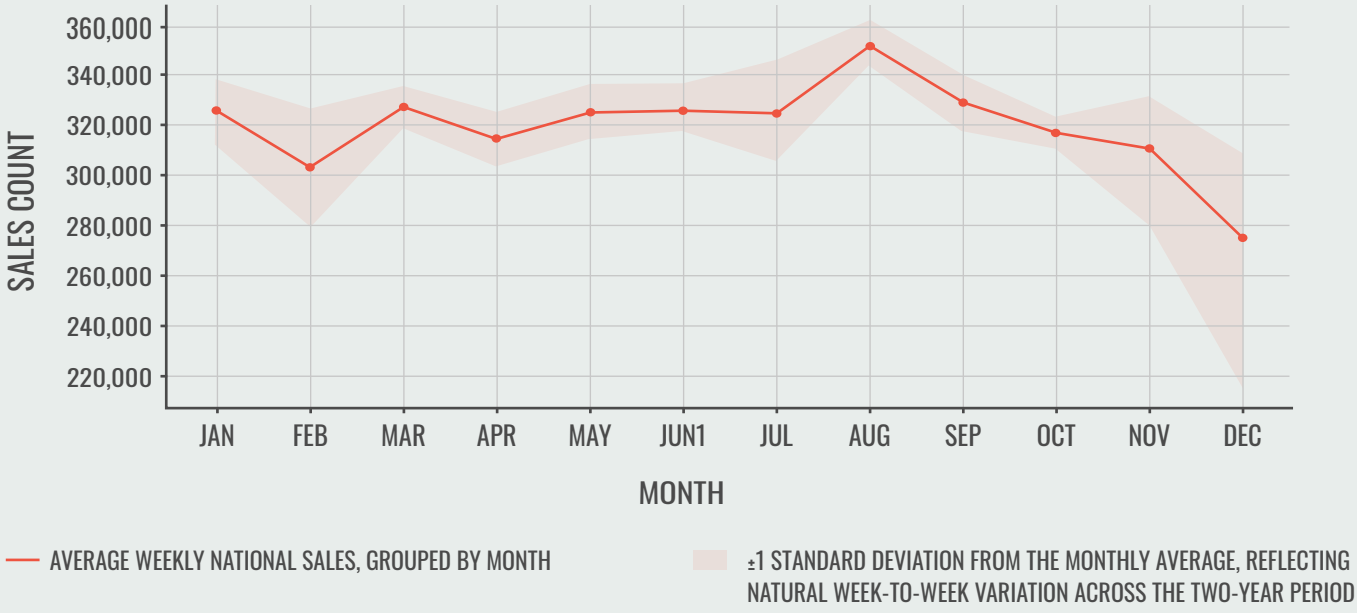
Post-Period (Observation Window)

Decide whether to measure lift only during exposure or include a tail period to capture delayed effects. Include a wash-out buffer if you expect lingering media influence that could contaminate post-test analysis. Pre-specify whether post-period sales will be (1) included in the primary lift calculation, (2) analyzed separately as a decay effect, or (3) excluded entirely as a buffer period. This prevents post-hoc decisions that could bias results.

Seasonality Alignment

Ensure that the pre-period and test period reflect comparable seasonal patterns. Avoid scheduling the experiment so that only one period includes major events like holidays, back-to-school, or Black Friday, unless both periods are expected to show similar seasonal dynamics. Misaligned timing can introduce artificial differences that may be mistaken for treatment effects. Marketing mix models can be used to create seasonally adjusted conversions in situations where bridging a "seasonal change" is unavoidable.

NATIONAL WEEKLY SALES TREND ACROSS TWO YEARS SHOWING SEASONALITY



▲ This line chart shows total weekly transactions across the U.S., grouped by month. While week-to-week variation is modest, a consistent seasonal pattern emerges: slower sales in early January, surges in mid-spring and summer, and a notable decline beginning around Thanksgiving that continues through Christmas. This trend is based on real national sales data for a Fortune 500 company and underscores the importance of pre-period normalization when designing geographic RCTs to control for seasonality that could otherwise bias treatment effect estimates.

GEOGRAPHIC CLUSTER SPECIFICATION

Unit Definition

In our example, we have determined to use DMAs as the cluster unit. They are mutually exclusive, cover the full U.S. population, and reflect media-buying logic with minimal spillover across boundaries. Their scale makes them well-suited for randomization and aggregate KPI tracking.

Sampling Frame

Use all 210 US DMAs rather than a subset to ensure generalizability to national campaigns. Full-market inclusion maximizes statistical power and captures regional heterogeneity, while limited-DMA designs introduce selection bias and restrict external validity of results.

Assignment Probability

Apply equal probability randomization (typically 50/50 for two arms, or 33/33/33 for three arms, etc.) across all included DMAs. This ensures each geographic unit has the same chance of receiving treatment, preserving the statistical properties needed for unbiased causal inference.

Exclusion Criteria

Define up front which DMAs may be excluded, and why. Valid reasons include things like low store count, data sparsity, regulatory constraints, or known structural anomalies. Aim to keep the full set of 210 DMAs unless exclusions are methodologically necessary. Excluding DMAs simply because they are large, account for a large portion of sales, or are competitive are not ideal exclusion criteria, as their inclusion inherently makes the results of the experiment more generalizable to real-world conditions. Forcing them into the test group is also a bad idea, moving out of the realm of being a randomized trial and into quasi-experiment territory.

Minimum Cluster Size

Set a clear inclusion threshold (e.g., “DMAs with fewer than 500 transactions per week may be excluded”) to avoid excess variance from underpowered regions. Document the rationale in the design phase to avoid ad hoc decisions post hoc.

Spillover/Contamination Safeguards

To prevent treatment effects from leaking into control DMAs, implement geo-targeting controls (e.g., frequency caps, DMA-level pacing). Set up monitoring protocols to track media delivery and investigate outliers.

OUTCOME DATA REQUIREMENTS

DATA SOURCE AND FORMAT

The experiment requires comprehensive sales (or other KPI) data at the geographic level for both historical power analysis (ideally 2 years) and the test period. Two primary sources are available:

First-Party CRM Data (Preferred):

- Extract daily or weekly transaction counts or sales volumes by ZIP code (daily preferred)
- Requires: buyer ZIP code and transaction date (daily or weekly) for aggregated sales counts
- Provides maximum flexibility and granularity
- No additional licensing costs

Third-Party Syndicated Data:

- Sources: NielsenIQ, Circana (formerly IRI), or similar providers
- Often pre-aggregated to DMA level
- Weekly granularity typical (daily may be cost-prohibitive but provides greater statistical power)
- Necessary for CPG and other industries without direct sales data

KPI DATA STRUCTURE

Outcome data for a geo test is typically simple, with three columns: date, sales count (or volume), and geography (ZIP code or DMA). If the advertiser has comprehensive first-party transaction data with ZIP codes in the customer record, outcomes should be aggregated by ZIP. If using a third-party panel such as NielsenIQ or Circana, DMA-level aggregation is sufficient. ZIP codes can be rolled up to DMAs using standard match tables from Nielsen or other providers.

While sales is the preferred outcome metric, other KPIs can be substituted in this structure, such as web form completions, brand search volume (from Google Trends, reported by DMA), or foot traffic (via third-party mobility data providers). Metrics can be reported daily or weekly. Sales can be expressed as counts or continuous dollar values.

EXAMPLE SALES DATA

DATE	SALES	ZIPCODE
7/28/2024	18	99353
7/28/2024	16	99352
7/28/2024	6	99350
7/28/2024	24	99338
7/28/2024	2	99337
7/28/2024	8	99336
7/28/2024	16	99324
7/28/2024	24	99323
7/28/2024	5	99320
7/28/2024	3	99301
7/28/2024	1	99163
7/28/2024	28	99141

DATA PREPARATION

For first-party data:

1. Extract transactions with valid ZIP codes and dates
2. Aggregate to daily or weekly counts/volumes by ZIP
3. Apply ZIP-to-DMA mapping table (available from Nielsen or other providers)
4. Sum to DMA-week or DMA-day level

For third-party data:

1. Verify that DMA definitions match media targeting parameters
2. Confirm data completeness and reporting lag
3. Align time periods with experimental design

METRIC SELECTION

Choose between transaction counts (categorical) or sales volume (continuous):

- **Counts:** More stable, less affected by outliers, simpler statistical properties
- **Volume:** Captures full business impact but may require transformation (log, square root) if highly skewed

Document this choice as it affects power calculations and analysis methods.

WHERE POSSIBLE, USE DAILY KPI AGGREGATIONS, WHICH PROVIDE MORE STATISTICAL POWER THAN WEEKLY. AVOID MONTHLY ROLL-UPS.

FORMULATE HYPOTHESIS

A good experimental hypothesis is a clear, testable statement about the expected causal effect of a treatment. It should specify:

- The direction and magnitude of expected impact (e.g., a lift of at least X%)
- The outcome metric (e.g., sales revenue)
- The comparison groups (treatment vs. control DMAs)
- The timeframe
- The confidence level required for the test

The hypothesis makes explicit what you believe the intervention will accomplish and provides stakeholders with a shared understanding of success criteria.

UNDERSTANDING HYPOTHESIS TESTING

Statistical testing doesn't "prove" a hypothesis true. Instead, it attempts to reject the opposite – the null hypothesis. The null hypothesis typically states there is no effect or that the effect is less than the pre-specified minimum. If the data are sufficiently inconsistent with the null hypothesis at your chosen confidence level, you reject it in favor of your research hypothesis.

Example Based on Business Objectives:
For the objective: "Measure iROAS of Meta as a channel partner and renegotiate or reallocate if results are unfavorable," the following hypothesis and null hypothesis could apply:

Research Hypothesis (H_1):
Meta advertising will generate a statistically significant lift in sales revenue of at least 3% in treatment DMAs compared to control DMAs during the 5-week test period, measured at the 95% confidence level.

Null Hypothesis (H_0):
Meta advertising will not generate a statistically significant lift in sales revenue of at least 3% in treatment DMAs compared to control DMAs during the 5-week test period at the 95% confidence level.

If we reject the null hypothesis, we have evidence supporting Meta's effectiveness. If we fail to reject it, either the true effect is smaller than 3%, the test lacked sufficient power, or Meta genuinely doesn't drive meaningful incremental sales at current spend levels.

Note that "failing to reject" the null hypothesis doesn't prove Meta ineffective; it may indicate the need for a longer test, higher spend, or acceptance that the channel drives modest but still positive returns below our detection threshold. Likewise, rejecting the null doesn't "prove" our hypothesis true; it provides evidence that the observed effect is unlikely to have occurred by chance alone. At a 95% confidence level, there's a 5% probability of incorrectly rejecting a true null hypothesis (Type I error). While we cannot prove hypotheses through statistical testing, we can accumulate evidence that makes certain conclusions more or less plausible given our data and assumptions.

RANDOMIZATION LOGIC

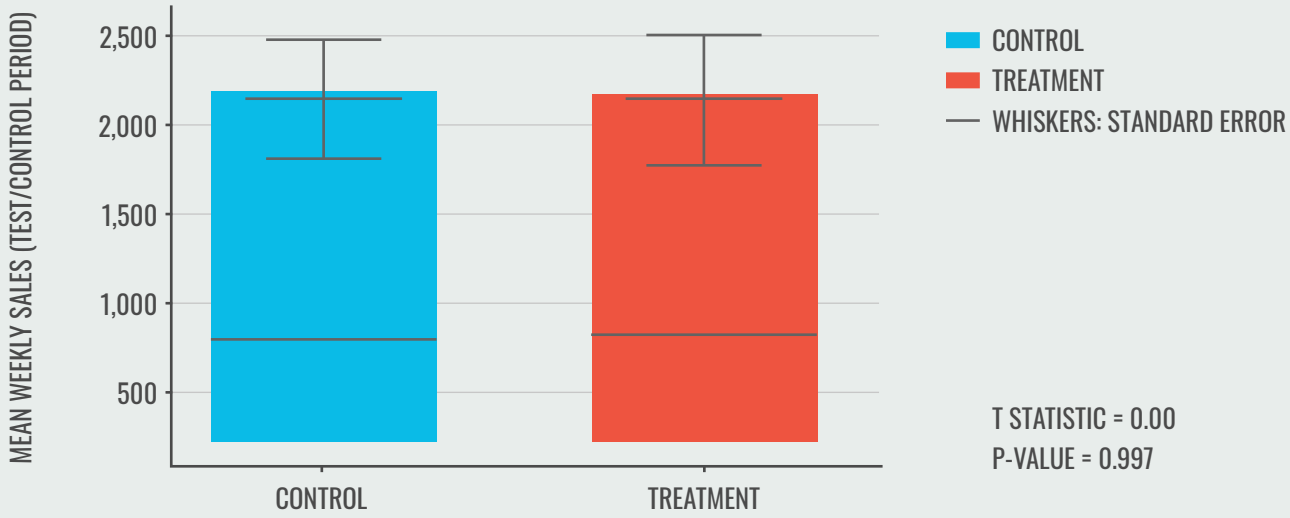
SIMPLE RANDOMIZATION AS THE DEFAULT

At its core, randomization is straightforward: we flip a fair coin (metaphorically) for each DMA to determine whether it receives treatment or serves as control. This simple approach, assigning each of the 210 DMAs to treatment or control with equal probability, forms the foundation of our experimental design.

(To see code for how it would work in practice, see Appendix A, Example 1: Simple Randomization for Geographic RCT.)

This approach might seem too simple – won't random chance sometimes create imbalanced groups? The key insight is that with 210 units and our comprehensive power analysis methodology, simple randomization works remarkably well , particularly in the multi-armed, stepped design dubbed Rolling Thunder.

ROLLING THUNDER BALANCE CHECK (6 GROUPS, 4 WEEKS EACH)



▲ This visualization demonstrates the balance achieved using the Rolling Thunder experimental framework with six test groups (A-F) and one control group during a 9-week experimental window, where the design enhances the control group by incorporating pre-treatment observations from test groups B through F before their respective 4-week treatment periods begin. With 150 DMAs randomly assigned across seven groups, the enhanced Control group achieved a mean of 2,136 sales per week (standard error = 335) compared to the Test group's mean of 2,134 sales per week (standard error = 368), with a two-sample t-test yielding t statistic = 0.03, p-value = 0.997, confirming no statistically significant difference between groups (<1% difference in means). The custom boxplot format displays the interquartile range (box), median (horizontal line within box), mean (thick horizontal line), and whiskers extending to mean ± 1 standard error, with no outliers shown. These values from a Fortune 500 company's actual sales data demonstrate how the Rolling Thunder framework naturally improves balance through random assignment without stratification or re-weighting, supporting robust causal inference while maintaining simplicity, with DMA-level normalization in the analysis phase further ensuring treatment effects aren't confounded by residual baseline differences.

Here's why:

OUR SIMULATION-BASED VALIDATION PROCESS

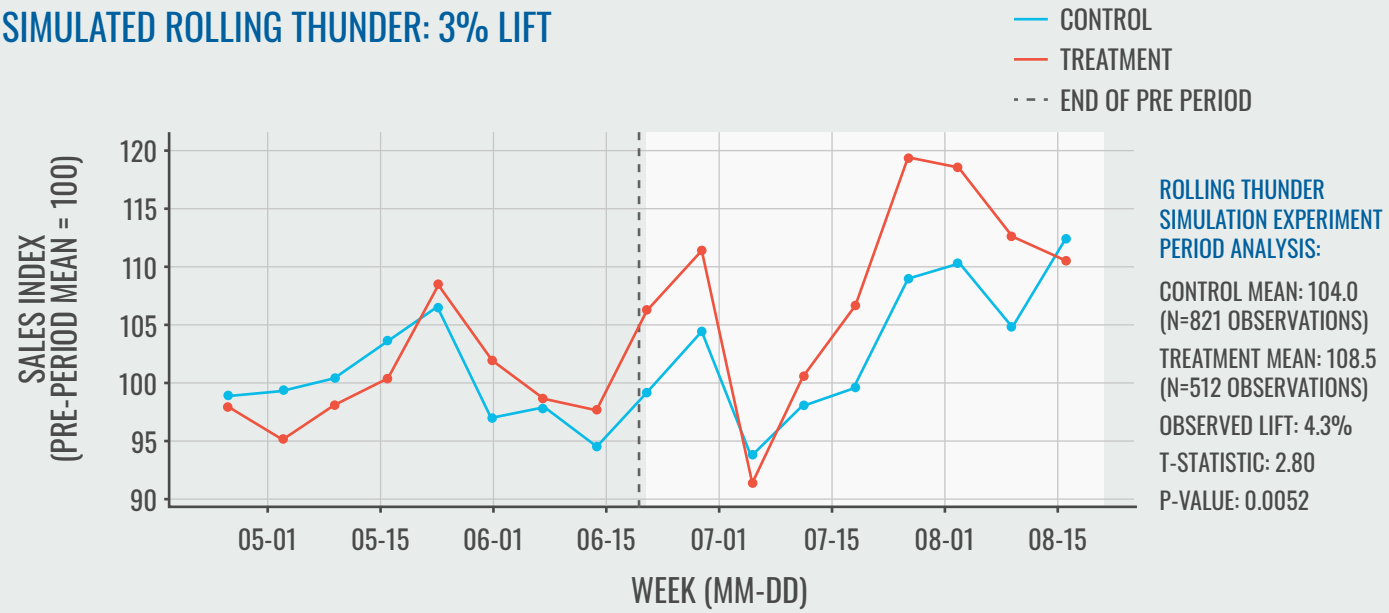
We run thousands of Monte Carlo simulations using ideally two years of historical weekly sales data. Each simulation:

1. **Randomly samples multiple test windows** from the historical period, preserving the calendar structure of pre- and post-periods. This ensures we test across different seasonal patterns and market conditions.
2. **Randomly assigns DMAs** to treatment and control groups using simple equal odds randomization, exactly as we would in the actual experiment.
3. **Normalizes each DMA to its own pre-period trend** using an 8-week baseline. This normalization process is crucial: it reduces baseline heterogeneity by accounting for each DMA's unique growth trajectory, seasonal patterns, and market dynamics before the test begins.
4. **Simulates a treatment effect** by artificially inflating sales in treatment DMAs during the test period by the hypothesized effect size (e.g., 3%, 5%).

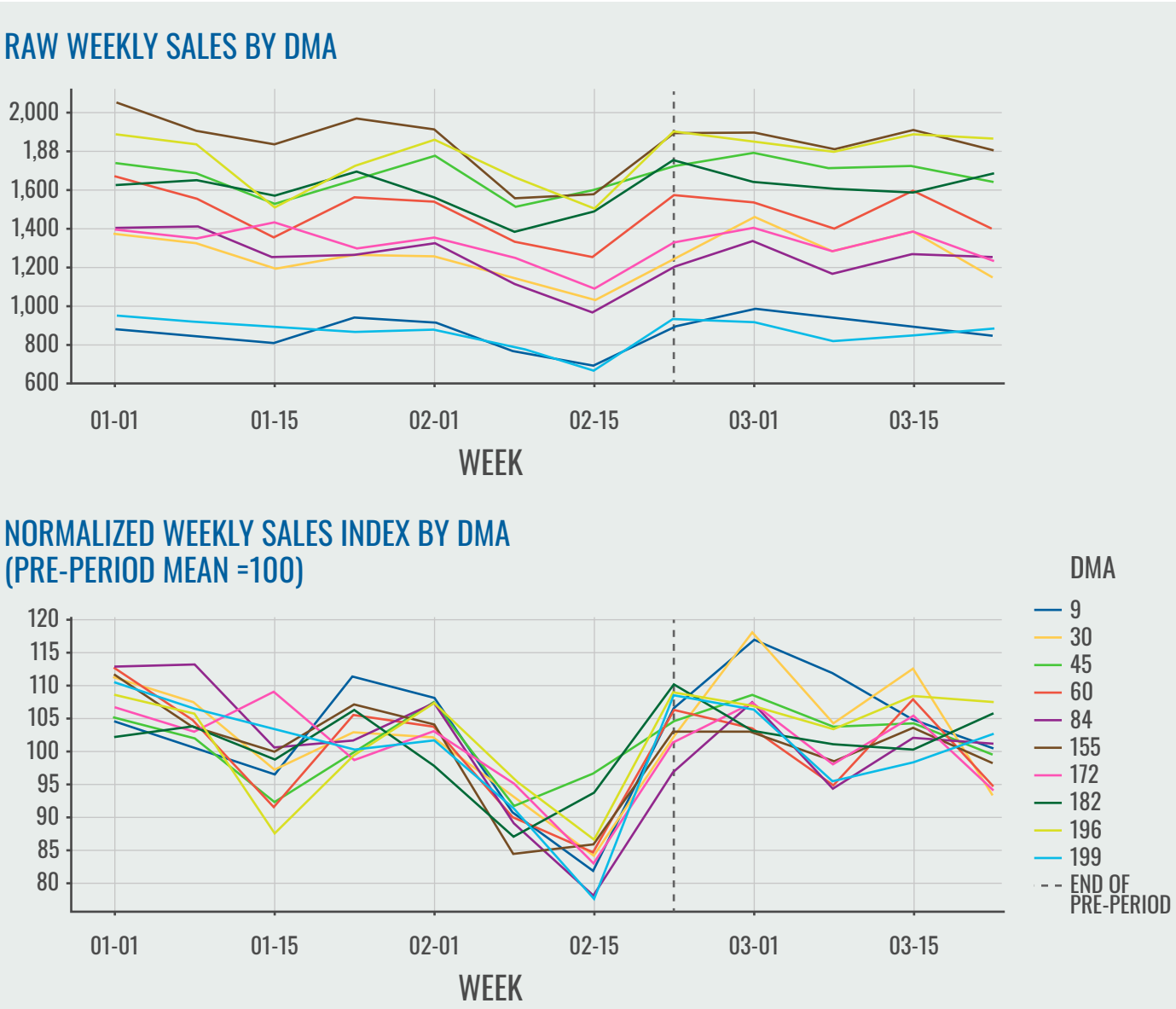
5. **Estimates treatment effects** using the planned analysis method (typically DMA-level t-test or ANCOVA).
6. **Tests for statistical significance** at the pre-specified confidence level (e.g., 95%).

This process is repeated hundreds or thousands of times with different test windows and treatment assignments. If more than 85% of simulations consistently show high power (P-val < 0.05) to detect target effects, we have empirical evidence that chance imbalances from simple randomization don't systematically undermine our design.

The pre-period normalization is particularly important. By normalizing each DMA to its own 8-week trend before treatment, we are effectively controlling for many potential confounders without having to explicitly model them. The Rolling Thunder design further enhances balance by incorporating the pre-treatment weeks from each test group into the control pool. This staggered entry means that test groups B through F contribute their early weeks as additional control observations, naturally increasing the effective control sample size and improving the representativeness of the control group across different time periods. This addresses the practical concern that with only 210 units, chance alone might produce treatment and control groups that differed systematically on important variables.



▲ **Simulated Rolling Thunder Results: Treatment vs. Control Trend.** This figure presents the outcome of a power-analysis simulation of a geo-experiment using the Rolling Thunder framework with historical sales data from a real Fortune 500 company. Six test groups (A-F) and one control group were created through random assignment of 150 DMAs, with each test group receiving treatment for 4 consecutive weeks starting in staggered fashion (Group A: weeks 1-4, Group B: weeks 2-5, etc.). The Rolling Thunder design enhances the control group by incorporating pre-treatment observations from test groups B through F before their respective treatment periods begin, naturally increasing the effective control sample size and improving balance. A 3% lift was artificially applied to each test group during its treatment period, following an 8-week pre-period used for normalization to each DMA's baseline (index = 100). The normalized weekly sales index shows clear divergence between the aggregated Treatment and enhanced Control groups during the 9-week experiment period. Statistical analysis using a two-sample t-test yields t = 2.16, p = 0.0363, confirming detection of the simulated lift at the 95% confidence level. The underlying sales data and randomization are real; only the treatment lift was simulated as part of a power analysis to assess the Rolling Thunder framework's statistical sensitivity. This would be one of thousands of simulations in the power analysis." This demonstrates that the combination of control enhancement through staggered entry and DMA-level normalization provides robust detection capability for modest treatment effects without requiring stratification or other complex balancing techniques.



▲ **Pre-period Normalization of Weekly Sales by DMA** These line plots show raw and normalized weekly sales across 10 example DMAs. The first panel (Raw Weekly Sales) reveals large differences in baseline market size and week-to-week variance, obscuring comparability across regions. The second panel (Normalized Weekly Sales Index) adjusts each DMA to its own pre-period mean (set to 100), enabling meaningful comparisons of relative change.

Summary statistics confirm the effect of normalization: raw sales had a standard deviation of 263 and a coefficient of variation (CV) of 0.165, whereas normalized sales had a standard deviation of 4.27 and a CV of 0.042. The sales data shown are actual transaction records from a Fortune 500 compay. The balance shown between test and control cells – after normalization using each DMA's pre-period trend – is genuine, not synthetic. This example demonstrates why stratification is typically unnecessary when using all 210 DMAs with this normalization method.

ALTERNATIVE RANDOMIZATION APPROACHES

Our simulation analyses demonstrate that the Rolling Thunder design inherently improves balance without requiring other such complex techniques. By incorporating the pre-treatment periods from all test groups (except the first) into the control group, this approach naturally enhances balance by leveraging observations from the same DMAs that will later receive treatment. The enhanced control group contains data from nearly all DMAs, more accurately reflecting the full population variance and making the framework robust against initial randomization imbalances. This natural balancing, combined with DMA-level normalization to pre-period means, typically provides sufficient statistical power for detecting modest lift effects. Certain situations may still benefit from more sophisticated approaches, such as these:

STRATIFIED RANDOMIZATION

Groups DMAs by key characteristics before randomizing within strata. (For sample code, see Appendix A, Example 12: Stratified Randomization.)

This guarantees balance on stratification variables but requires careful selection of stratifying covariates. Best when you can identify 2-3 clearly important variables (i.e., with strong correlation to KPI) and have sufficient DMAs within each stratum.

Example: Stratify by sales quartile to ensure treatment and control arms have equal representation of high-, medium-, and low-volume markets.

MATCHED-PAIR RANDOMIZATION

Pairs similar DMAs based on multiple characteristics (e.g., past sales trends, populations, competitive mix), then randomly assigns one from each pair to treatment. Particularly valuable when sample sizes are smaller or effect sizes are subtle.

IMPORTANT DISTINCTION: MATCHED-PAIR RANDOMIZATION VS. MATCHED-MARKET TESTING

While the names sound similar, matched-pair randomization is fundamentally different from matched-market testing. In matched-market testing (a quasi-experimental method), you hand-pick specific control markets that seem similar to your test markets: there's no randomization. In matched-pair randomization, you still use random assignment, but you do it within pairs of similar markets, and large numbers of them. This preserves the causal inference benefits of randomization while improving balance.

Implementation:

First, calculate a similarity score between all possible DMA pairs based on key metrics like sales history, population, and demographics. This can be done using statistical distance measures that quantify how "far apart" two DMAs are across multiple dimensions. Then, match each DMA with its most similar partner. Finally, use a randomization algorithm within each pair to assign one to treatment and one to control.

COVARIATE-CONSTRAINED RANDOMIZATION (RE-RANDOMIZATION)

Generates multiple random assignments (e.g., 10,000 draws) and selects one meeting pre-specified balance criteria. Provides precise control over balance without sacrificing randomization principles.[6]

[6] Rerandomization To Improve Covariate Balance In Experiments" (Morgan, Rubin, 2012)

This is not a form of "p-hacking" (a bad practice, in this case repeatedly analyzing data until you find a significant result) because you're not looking at outcomes, you're only checking whether the random assignment created balanced groups before the experiment even starts. As long as your balance criteria are defined before seeing any results, this technique maintains the integrity of the randomization.

Balance assessment:

Use standardized mean differences (SMD) with threshold of absolute SMD < 0.1 for core covariates like pre-period sales and DMA population.

POST-STRATIFICATION AND COVARIATE ADJUSTMENT

Proceeds with initial randomization but adjusts for imbalances in the statistical analysis phase through regression or ANCOVA. Doesn't change assignment but can recover power lost to imbalance.

These techniques become particularly valuable when:

- Working with fewer than 210 DMAs (e.g., regional tests)
- Unable to run extensive historical simulations
- Stakeholders require demonstrable balance for confidence
- Effect sizes are expected to be small (<2%)

IMPORTANT DISTINCTION: RCT ENHANCEMENTS VS. QUASI-EXPERIMENTAL METHODS

The approaches above are all enhancements within the RCT framework – they preserve randomization while improving efficiency. This is fundamentally different from quasi-experimental approaches like:

- **Propensity Score Methods** in observational studies that reweight based on estimated treatment probability
- **Synthetic Control Methods** that construct artificial control units from weighted combinations and are subject to over-fitting

TREATMENT INTENSITY ("DOSAGE")

QUANTIFICATION

Define the planned treatment dose in media terms: impressions, GRPs, TRPs, or spend per capita. Be specific and use a KPI that maps cleanly to lift models.

RATIONALE

Confirm that this dosage is expected to produce a lift equal to or greater than the expected effect size. Link back to prior experiments or planning models to justify that the level of spend can drive a detectable effect.

MONITORING PLAN

Set up real-time dashboards or pacing reports to ensure that delivery is on track. Allow for mid-flight correction without altering randomization, especially in high-stakes or high-budget tests.

TREATMENT-DELIVERY VERIFICATION

Establishing robust monitoring systems before launch ensures that the intended media exposure actually reaches treatment DMAs while staying out of control DMAs, a critical requirement for valid causal inference.

DATA SOURCE

Specify the systems (e.g., DSP logs, third-party verification platforms, media vendor reports) that will confirm media delivery at the geographic unit level.

COMPLIANCE METRICS

Define acceptable delivery bounds, e.g., "A DMA is compliant if it receives ≥90% of planned impressions within ±10% of schedule." Track and report compliance systematically for each treatment DMA.

DATA QUALITY & LATENCY RULES

LATENCY TOLERANCE

Specify acceptable reporting lag, e.g., "All sales must be posted to the warehouse within five days of transaction date." Define a cut-off date for primary analysis to allow for complete data.

DATA AUDITS

Schedule at least one interim extract during the test to verify record counts, completeness, and DMA-level mapping. Use this as a QA step to confirm integrity before final analysis.

BUSINESS SAFEGUARDS

MINIMUM CONTROLS

If relevant (e.g., in a cessation test), declare a minimum media presence in control DMAs to avoid business disruption or brand risk.

EARLY-STOP CRITERIA

If conditional stopping is allowed, e.g., to prevent severe loss of sales in a cessation test, pre-register the interim check schedule and thresholds. For example: "Interim analysis after 2 weeks; stop test if sales decline exceeds 15% with 98.5% confidence."

Note that checking results early requires more stringent significance thresholds to maintain overall experiment integrity – if you peek at results multiple times, you increase the chance of false conclusions. Common approaches include requiring $p < 0.015$ for early stops (versus $p < 0.05$ for final analysis) or using sequential testing methods that adjust thresholds based on the number of looks.

Define governance clearly: who has authority to stop the test, what evidence is required, and what happens to the analysis if early stopping occurs.

OPTIONAL: MULTIPLE OUTCOMES & SUBGROUPS

While we strongly recommend focusing on a single primary KPI to maintain clarity and statistical rigor, some tests may require secondary analyses. The critical principle: any additional outcomes must be specified during the design phase, not added post hoc after seeing results.

SECONDARY KPIS

If you must track additional metrics (e.g., transactions, average order value, new-customer rate), pre-declare whether p-values will be corrected for multiple comparisons (e.g., Holm-Bonferroni) or considered purely exploratory. Remember that each additional hypothesis test increases the risk of false positives.

SUBGROUP ANALYSIS

Similarly, any planned subgroup analyses (e.g., high- vs. low-income DMAs, urban vs. rural) should be specified at the design phase. These will be tested using interaction terms in the primary model, not as separate experiments. Be cautious: subgroup analyses are often underpowered and can lead to spurious findings.

CONDUCT POWER ANALYSIS

Power analysis is a critical step in designing a geographic RCT. It evaluates whether the experiment design, given its parameters, can reliably detect a meaningful effect size. Rather than relying solely on theoretical variance assumptions, we use a simulation-based power analysis grounded in real-world sales data. This approach provides an empirical check on whether the design is likely to produce statistically significant results, and helps quantify the test's sensitivity under realistic conditions.

HISTORICAL DATA REQUIREMENTS FOR SIMULATION

The simulation requires the same data format as the actual experiment. (For a code example, see Appendix A, Example 2: Power Analysis Data Structure.)

For first-party data, this means:

1. Two years of historical ZIP-level transactions
2. Same aggregation logic as planned for test period
3. Same ZIP-to-DMA mapping to ensure consistency

For syndicated data:

1. Negotiate access to historical data (ideally 2 years)
2. Verify no methodology changes during this period
3. Account for any markets with incomplete history

The power simulation will use this exact data structure, preserving real-world variance patterns, seasonality, and DMA-specific trends that theoretical calculations would miss.

SIMULATION-BASED POWER ESTIMATION

We simulate the experiment using ideally two years of historical weekly sales data to reflect true patterns of variance, seasonality, and DMA-level idiosyncrasies. The simulation process mirrors the structure of the planned experiment:

Step-by-Step Simulation Process:

(For sample code, see Appendix A, Example 13: Power Simulation Framework for detailed implementation.)

1. **Randomly sample multiple possible test windows** from the historical period. For a planned 6-week test, we would sample dozens of different 6-week windows across the two years, preserving the calendar structure of pre- and post-periods. This captures different seasonal contexts – a test in Q4 may behave differently than Q2.
2. **Vary the treatment duration** systematically. Run parallel simulations at 3, 4, 5, 6, 8, 10, and 12 weeks. While 4-6 weeks is often the sweet spot for balancing cost and statistical power, only simulation reveals the actual trade-offs for your specific sales patterns.
3. **Apply the DMA normalization process** that mirrors our actual analysis. For each simulation:

POWER GRID: P-VALUE ESTIMATES FOR MULTI-ARMED ROLLING THUNDER TEST BY SALES/MONTH FOR VARIOUS EFFECT SIZES

MONTHLY CONVERSIONS	MINIMUM DETECTABLE EFFECT SIZE						
	0.1%	0.2%	0.5%	1.0%	5.0%	10.0%	20.0%
500	47.4%	52.7%	44.0%	56.6%	37.0%	61.8%	35.4%
5,000	60.1%	48.2%	33.1%	51.4%	55.2%	21.6%	0.0%
10,000	41.4%	50.6%	43.6%	52.5%	25.1 %	0.3%	0.0%
25,000	50.2%	39.8%	58.9%	45.1%	12.2%	0.0%	0.0%
50,000	57.9%	57.1%	43.9%	42.6%	0.1%	0.0%	0.0%
100,000	49.9%	47.0%	44.6%	41.9%	0.0%	0.0%	0.0%
200,000	47.9%	41.5%	41.7%	32.3%	0.0%	0.0%	0.0%
500,000	51.1%	44.7%	23.4%	5.3%	0.0%	0.0%	0.0%
1,000,000	52.8%	44.9%	12.1%	0.5%	0.0%	0.0%	0.0%
2,500,000	57.0%	23.7%	1.1%	0.0%	0.0%	0.0%	0.0%

▲ **Cell values show p-values from statistical tests at each combination of monthly conversions and minimum detectable effect size.** This example shows the following assumed parameters: control fraction of 20%; 10 weeks total experiment time; 7 groups across all 210 DMAs; conversions distributed in proportion to population. Green cells: Statistically significant at the 95% confidence level ($p < 0.05$). Orange to red cells: Increasingly non-significant results ($p \geq 0.05$), indicating lower power to detect the effect.

- Calculate each DMA's average weekly sales during its 8-week pre-period
- Compute the week-over-week growth rate during this baseline
- Project expected sales during the test period based on this trend
- Express actual test-period sales as an index relative to this projection. This normalization is crucial: it transforms raw sales (which vary greatly by DMA size) into comparable lift indices, dramatically reducing variance.

4. **Randomly assign DMAs to treatment and control groups** using the same randomization logic planned for the real experiment (typically equal odds of assignment to all experiment arms).
5. **Simulate a treatment effect** by inflating normalized sales values in the treatment group during the test period by the hypothesized effect size (e.g., 3%, 5%). This creates the "signal" we're trying to detect against the "noise" of natural variation.
6. **Estimate treatment effects using the planned analysis method** (e.g., DMA-level t-test on the normalized values).
7. **Test for statistical significance** at the pre-specified confidence level (e.g., 95%).
8. **Repeat this process thousands of times**, with different test windows, treatment assignments, and durations to build a robust empirical distribution of power outcomes.

KEY SIMULATION OUTPUTS

The Monte Carlo simulation yields several critical insights:

- **Power curves by duration:** Shows probability of detecting effects at 3%, 5%, 7% lift for each test length
- **Minimum detectable effect (MDE):** The smallest lift where power exceeds 80%
- **Optimal test configuration:** The duration that minimizes MDE while keeping the test practical

For example, simulations might reveal:

- 3-week test: 80% power only at 7%+ lift (too insensitive)
- 5-week test: 80% power at 3.5% lift (good balance)

- 8-week test: 80% power at 2.8% lift (marginal improvement for 60% more media cost)

ESTIMATING AND APPLYING INTRACLASS CORRELATION COEFFICIENT (ICC)

Within the simulation process, we also estimate the Intraclass Correlation Coefficient (ICC). This measures how similar sales values are within each DMA over time relative to between DMAs, and informs how much clustering is reducing the effective number of observations.[7]

Understanding ICC Impact:

- **Low ICC (< 0.05):** DMAs behave relatively independently week-to-week. Each week adds substantial information.
- **High ICC (> 0.20):** Strong within-DMA correlation. Adding weeks provides diminishing returns.

A high ICC indicates that data within each DMA are highly correlated – meaning that more weeks (or switching to daily data) are needed to compensate for the lack of independence. Because we treat the number of DMAs as fixed (typically all 210), we must adjust for ICC through test duration and data granularity:

Practical Adjustments for High ICC:

1. **Extend the test window** to observe more time points. If ICC = 0.15, you might need 6 weeks instead of 4 to achieve target power.
2. **Switch to daily data**, if day-to-day variation adds usable signal. Daily data often shows lower ICC than weekly aggregates, effectively increasing sample size.
3. **Run parallel simulations** with different granularities and durations to identify the most power-efficient combination. Sometimes 4 weeks of daily data outperforms 6 weeks of weekly data.

(For a code example, see Appendix A, Example 3: Estimating And Applying Intraclass Correlation Coefficient).

SETTING POWER TARGETS AND MAKING DECISIONS

STATISTICAL POWER TARGET

Statistical power is the probability that the experiment will detect a real effect if one exists. A typical power target is 80%, though 90% may be preferred for high-stakes tests. This target should be pre-committed and documented in the protocol to avoid post hoc reinterpretation.

DECISION FRAMEWORK

Once the simulation yields a power curve, identify the design configuration (duration, data granularity, MDE) that meets or exceeds the power target. The goal is to ensure that if the treatment works as hypothesized, it will be reliably detected, given the noise and clustering in real-world data.

Key decision points:

1. If power is too low at business-relevant effect sizes:
   - Extend test duration (most common solution)
   - Increase media weight to drive larger effects
   - Broaden KPI scope (e.g., full brand line vs. single product)
   - Switch to daily data if available
2. If power is very high (>95%):
   - Consider shortening the test to save budget
   - Test more subtle tactics or lower spend levels
   - Run multiple sequential tests instead of one long test
3. Document the final configuration:
   - "5-week test achieves 83% power to detect 3% lift"
   - "Daily data improves MDE from 3.5% to 2.8% vs. weekly"
   - "Extending beyond 6 weeks provides minimal power gain"

[7] Design and Analysis of Cluster Randomized Trials" (Crespi, 2019)

LOCK THE DESIGN FRAMEWORK

Once power analysis confirms feasibility, document every parameter in a formal experiment design document. This should be circulated to all stakeholders (advertiser, agency, media partners, analytics team) before any media is trafficked.

ESSENTIAL COMPONENTS

GEOGRAPHIC RCT DESIGN SPECIFICATION EXAMPLE

Business Question	Does Social Media drive ≥3% lift in sales revenue?
Decision Rule	If lift ≥3%, increase budget 20%. If <3%, reallocate 50%.
Primary KPI	Weekly sales revenue by DMA (from CRM, ZIP→DMA mapping)
Experimental Unit	210 US DMAs (Nielsen 2024 definitions)
Design Type	Multi-armed stepped trial ("Rolling Thunder")
Randomization	Simple random assignment, equal probability
Random Seed	42 (stored in repo: geoRCT_Q2_2025_assignments.csv)
Pre-Period	8 weeks (Mar 1 - Apr 25, 2025)
Treatment Period	5 weeks (Apr 26 - May 30, 2025)
Post-Period	2 weeks (May 31 - Jun 13, 2025) [washout buffer]
Treatment Dosage	10M impressions per week @ $10 CPM (≈$500,000 total)
Media Channels	Facebook, Instagram
Confidence Level	95% (α = 0.05)
Power Target	≥80%
Simulation Results	84% power to detect 3% lift (via 2 years historical data) MDE = 2.7% (minimum detectable effect at 80% power)
Normalization Method	8-week pre-period trend per DMA
Analysis Method	T-test on normalized sales indices
Compliance Threshold	Treatment DMAs must receive 90-110% of planned spend
Exclusions	None (all 210 DMAs included)
Secondary Analyses	None pre-specified
Governance	CMO approves changes; Analytics owns methodology

REQUIRED ATTACHMENTS:

1. **DMA Assignment File**
   (geoRCT_Q2_2025_assignments.csv):
2. **Power Analysis Output**
   (power_simulation_results.html):
   - Power curves by effect size and duration
   - ICC estimates and sensitivity analysis
   - Simulation code for reproducibility
3. **Media Trafficking Instructions**
   (media_setup_guide.pdf):
   - Platform-specific targeting lists
   - Creative rotation rules
   - Pacing and budget allocation by week
4. **Data QA Checklist**
   (pre_launch_qa.xlsx):
   - Verify ZIP→DMA mapping is current
   - Confirm sales data pipeline latency
   - Test compliance monitoring dashboard
   - Validate historical data completeness

Version Control:

Store all documents in a version-controlled repository (Git, SharePoint with versioning, etc.). Any changes after stakeholder sign-off require:

1. Written change request with rationale
2. Impact analysis on statistical power
3. Approval from both business and analytics leads
4. Updated version with change log

EXAMPLE RANDOMIZED DMA LIST

CONTROL	TALLAHASSEE-THOMASVILLE, YAKIMA-PASCO-RCHLND-KNNWCK, BAKERSFIELD, BALTIMORE, EUGENE, MINNEAPOLIS-SAINT PAUL, PEORIA-BLOOMINGTON, LITTLE ROCK-PINE BLUFF, BEND, OR, TRI-CITIES, TN-VA, WASHINGTON, DC-HAGRSTWN, SAN FRANCISCO-OAKLAND-SAN JOSE, MADISON, PARKERSBURG, HARTFORD-NEW HAVEN, BUTTE-BOZEMAN, SAINT LOUIS, SOUTH BEND-ELKHART, AMARILLO, ROANOKE-LYNCHBURG, FRESNO-VISALIA, SYRACUSE, BILLINGS, CHICO-REDDING, ALBUQUERQUE-SANTA FE, JUNEAU, TYLER-LONGVIEW (LFKN&NCGD)
GROUP A	QUINCY-HANNIBAL-KEOKUK, TULSA, JACKSON, TN, PORTLAND-AUBURN, UTICA, BATON ROUGE, COLUMBIA-JEFFERSON CITY, LIMA, COLORADO SPRINGS-PUEBLO, HOUSTON, ALPENA, TOLEDO, SAN DIEGO, MISSOULA, OTTUMWA-KIRKSVILLE, ALBANY-SCHENECTADY-TROY, SPRINGFIELD-HOLYOKE, HARRISONBURG, GREENSBORO-HIGH POINT-WINSTON SALEM, COLUMBUS-TUPELO-WEST POINT, COLUMBUS, GA, WICHITA FALLS-LAWTON, WAUSAU-RHINELANDER, WILKES BARRE-SCRANTON, HARRISBURG-LANCASTER-LEBANON-YORK, HUNTSVILLE-DECATUR-FLORENCE
GROUP B	MONROE-EL DORADO, ANCHORAGE, AUGUSTA, SEATTLE-TACOMA, MYRTLE BEACH-FLORENCE, MIAMI-FORT LAUDERDALE, DALLAS-FORT WORTH, LUBBOCK, LA CROSSE-EAU CLAIRE, CHICAGO, LOUISVILLE, ROCKFORD, ODESSA-MIDLAND, NEW YORK, ROCHESTER-MASON CITY-AUSTIN, CLEVELAND-AKRON, RENO, ELMIRA, HELENA, MILWAUKEE, MINOT-BISMARCK-DICKINSON, TWIN FALLS, BOSTON, LAFAYETTE, IN, BLUEFIELD-BECKLEY-OAK HILL, LINCOLN-HASTINGS-KEARNEY PLUS
GROUP C	MOBILE-PENSACOLA, JOHNSTOWN-ALTOONA, SPRINGFIELD, MO, TRAVERSE CITY-CADILLAC, NEW ORLEANS, LOS ANGELES, DULUTH-SUPERIOR, WEST PALM BEACH-FT PIERCE, NORTH PLATTE, JOPLIN-PITTSBURG, CLARKSBURG-WESTON, FORT WAYNE, WACO-TEMPLE-BRYAN, PRESQUE ISLE, GREENVILLE-SPARTANBURG-ASHEVILLE, DENVER, CHARLOTTE, SAN ANTONIO, PORTLAND, OR, FARGO-VALLEY CITY, AUSTIN, MONTEREY-SALINAS, OKLAHOMA CITY, CASPER-RIVERTON, TERRE HAUTE, IDAHO FALLS-POCATELLO
GROUP D	BILOXI-GULFPORT, LAKE CHARLES, YOUNGSTOWN, MANKATO, LAREDO, DES MOINES-AMES, WATERTOWN, OMAHA, MARQUETTE, JONESBORO, LANSING, JACKSONVILLE, PANAMA CITY, ATLANTA, BANGOR, CORPUS CHRISTI, DAYTON, PHILADELPHIA, CHEYENNE-SCOTTSBLUFF, PADUCAH-CAPE GIRARDEAU-HARRISBURG, GLENDIVE, ERIE, DETROIT, MEDFORD-KLAMATH FALLS, SHREVEPORT, KANSAS CITY
GROUP E	EUREKA, PHOENIX, GREAT FALLS, RAPID CITY, SHERMAN-ADA, SALT LAKE CITY, CHARLESTON-HUNTINGTON, BINGHAMTON, FLINT-SAGINAW-BAY CITY, COLUMBIA, SC, ABILENE-SWEETWATER, CHAMPAIGN-SPRINGFIELD-DECATUR, KNOXVILLE, SPOKANE, EL PASO, CHATTANOOGA, SIOUX FALLS (MITCHELL), GREEN BAY-APPLETON, SIOUX CITY, FAIRBANKS, CHARLESTON, SC, WHEELING-STEUBENVILLE, FORT SMITH-FAY-SPRNGDL, BIRMINGHAM, BOISE, ORLANDO-DAYTONA BEACH-MELBOURNE
GROUP F	HATTIESBURG-LAUREL, BURLINGTON-PLATTSBURGH, LAS VEGAS, ALEXANDRIA, LA, DOTHAN, MEMPHIS, RICHMOND-PETERSBURG, MONTGOMERY, BEAUMONT-PORT ARTHUR, WILMINGTON, BUFFALO, MERIDIAN, GRAND JUNCTION-MONTROSE, GREENWOOD-GREENVILLE, LEXINGTON, GRAND RAPIDS-KALAMAZOO-BATTLE CREEK, SAVANNAH, PITTSBURGH, CEDAR RAPIDS-WATERLOO-DUBUQUE, MACON, NASHVILLE, SANTA BARBARA-SAN MAR-SAN LUIS OBISPO, SACRAMENTO-STOCKTON-MODESTO, BOWLING GREEN, JACKSON, MS, INDIANAPOLIS
GROUP G	DAVENPORT-ROCK ISLAND-MOLINE, LAFAYETTE, LA, FORT MYERS-NAPLES, CINCINNATI, TUCSON-SIERRA VISTA, HARLINGEN-WESLACO-BROWNSVILLE, PROVIDENCE-NEW BEDFORD, GAINESVILLE, CHARLOTTESVILLE, TAMPA-ST PETERSBURG-SARASOTA, ALBANY, GA, RALEIGH-DURHAM, COLUMBUS, OH, TOPEKA, VICTORIA, HONOLULU, WICHITA-HUTCHINSON PLUS, SALISBURY, GREENVILLE-NEW BERN-WASHINGTON, SAN ANGELO, YUMA-EL CENTRO, NORFOLK-PORTSMOUTH-NEWPORT NEWS, ROCHESTER, NY, SAINT JOSEPH, ZANESVILLE, EVANSVILLE

EXECUTE THE EXPERIMENT

PRE-LAUNCH VERIFICATION

Before launching the test campaign run through this checklist:

Media Setup Verification:

- ☐ Treatment DMA lists uploaded to all platforms
- ☐ Control DMA exclusions properly configured, if necessary for a given medium (otherwise, pure holdout will suffice)
- ☐ Geo-targeting set to “exact” (no radius expansion)
- ☐ Frequency caps consistent across platforms
- ☐ Creative assets identical in all DMAs
- ☐ Budget distributed across Treatment DMAs proportional to pre-test KPI rate

Data Pipeline Verification:

- ☐ Historical sales data passes integrity checks
- ☐ ZIP→DMA mapping uses current DMA definitions consistent with the medium’s targeting
- ☐ Sales reporting latency documented and acceptable
- ☐ Compliance monitoring dashboard live
- ☐ Backup data extraction plan in place

LAUNCH AND MONITOR

WEEK 1 - SOFT LAUNCH:

- Daily compliance checks: spend by DMA vs. plan
- Verify targeting accuracy
- Investigate any Control DMA with >0% spend
- Verify Treatment DMAs receiving impressions

WEEKS 2-5 - FULL FLIGHT:

Establish a monitoring cadence:

```
Monday AM  : Pull weekend delivery data
Monday PM  : Calculate compliance metrics by DMA
Tuesday AM : Flag DMAs outside 90-110% of plan
Tuesday PM : Adjust bids/budgets (not targeting)
Wednesday  : Spot-check creative rotation
Thursday   : Review week-to-date pacing
Friday     : Send compliance report to stakeholders
```

CRITICAL RULES:

1. **Never change DMA assignments** - This breaks randomization
2. **Document all adjustments** - Include timestamp and rationale
3. **Don't peek at results** - Avoid the temptation to check lift mid-flight unless conditional stopping was pre-established
4. **Maintain spend ratios** - If cutting budget is necessary, due to business conditions, cut proportionally across all DMAs

COMPLIANCE MONITORING

Track these metrics weekly:

METRIC	ON TARGET	YELLOW FLAG	RED FLAG
% TREATMENT DMAS COMPLIANT	95%+	90-94%	<90%
% CONTROL DMAS WITH SPILLOVER	<5%	5-10%	>10%
SPEND VARIANCE ACROSS TREATMENT	<20%	20-30%	>30%
PLATFORM DELIVERY DISCREPANCIES	<10%	10-20%	>20%

COMMON MID-FLIGHT ISSUES AND SOLUTIONS

Problem: Underspending in small DMAs

- Solution: Increase bids/spend, not targeting parameters
- Document markets affected and adjustment dates

Problem: Platform reporting discrepancies

For example, Facebook Ads Manager shows 1,200 conversions while Google Analytics shows 850 for the same campaign due to different attribution windows (view-through vs. click-only) and tracking methods.

- Solution: Designate one platform as "source of truth"
- Reconcile differences post-campaign

Problem: Competitive activity spike

- Solution: Continue as planned; document for post-analysis
- Do not add markets or extend test reactively

CONDUCT ANALYSIS

DATA PREPARATION

For CRM Data:

After the test period ends, follow the same data extraction process used for historical data.

(For sample code, see Appendix A, Example 4: SQL to Python – Extract Raw Transactions.)

Then apply ZIP-to-DMA mapping.

(For sample code, see Appendix A, Example 5: Apply ZIP to DMA Mapping and Aggregate.)

For Syndicated Data:

Request data pull with these specifications:

- Markets: All 210 US DMAs (or subset used in test)
- Metrics: [Total sales volume / unit sales / transactions]
- Time period: [Full date range including pre and post periods]
- Granularity: Daily (preferred, more statistical power) or weekly

Allow 5-7 business days after period close for transaction settlement and data processing, then freeze the dataset for analysis.

DATA QUALITY CHECKS

If using sales volume (continuous), check distribution of normalized values.

For count data, standard t-test typically suffices unless counts are very low (< 5 per DMA-week), in which case consider Poisson regression.

(For sample code, see Appendix A, Example 6: Data Quality Checks.)

PRIMARY ANALYSIS: T-TEST ON NORMALIZED VALUES

The core analysis follows these steps:

1. Calculate pre-period baselines for each DMA
2. Normalize test period sales to account for DMA heterogeneity
3. Compare normalized values between Treatment and Control
4. Test for statistical significance

(For sample code, see Appendix A, Example 7: Primary Analysis – t-test on Normalized Values.)

INTERPRETING RESULTS:

The coefficient on assignment Treatment represents the estimated lift. For example:

```
Estimate: 0.0342
Std. Error: 0.0156
t value: 2.19
p-value: 0.029
95% CI: [0.0036, 0.0648]
```

This indicates a 3.42% lift (p = 0.029), statistically significant at the 95% confidence level.

OPTIONAL ROBUSTNESS CHECKS

While randomized controlled trials (RCTs) provide the strongest basis for causal inference, additional diagnostic checks can help validate that the estimated treatment effect is not an artifact of pre-existing differences or instability in the data. These checks are not required in every case but may be useful for stakeholder reassurance or internal QA in higher-stakes tests.

For sample code, see Appendix A:

- Example 8 – Difference-in-Differences Cross-Check
- Example 9 – Leave-One-Out Sensitivity Analysis
- Example 10 – Pre-Period Balance Check

SUMMARY TABLE: DIAGNOSTIC CHECKS ON SIMULATED RCT RESULTS

CHECK	ESTIMATE	P-VALUE	INTERPRETATION
DIFFERENCE-IN-DIFFERENCES	0.97	-	CONFIRMS TREATMENT EFFECT REMAINS WHEN COMPARING PRE/POST TRENDS ACROSS GROUPS.
PRE-PERIOD BALANCE CHECK (T-TEST)	1.0	1.0000	NO SIGNIFICANT DIFFERENCES BEFORE TREATMENT; GROUPS WERE WELL BALANCED AT BASELINE.
LEAVE-ONE-OUT (MEAN EFFECT)	3.61		TREATMENT EFFECT IS STABLE ACROSS GEOGRAPHIES; NO SINGLE DMA DRIVES THE RESULT.

DESCRIPTION OF CHECKS

1. Difference-in-Differences (DiD)

Compares the change in outcome between Treatment and Control groups from pre- to post-period. This provides a cross-check that the estimated lift is not merely due to different group trajectories over time.

2. Pre-Period Balance Check

A simple t-test comparing average pre-treatment outcomes between Treatment and Control groups. A high p-value (e.g., >0.1) suggests the groups were balanced before treatment, reducing the risk of confounding.

3. Leave-One-Out Sensitivity

Runs the treatment effect estimation multiple times, each time excluding one DMA from the Treatment group. If the estimated lift remains stable across iterations, the result is not overly influenced by any single region.

DOCUMENTING RESULTS

Create a comprehensive results document:

```
GEO RCT RESULTS SUMMARY

Test Period: Apr 26 - May 30, 2025
PRIMARY RESULT:
Lift Estimate: +3.42% (95% CI: 0.36% to 6.48%)
P-value: 0.029
Statistical Significance: YES at 95% confidence
BUSINESS IMPACT:
- Test Spend: $3.2MM
- Incremental Revenue: $8.7MM
- iROAS: 2.72
- Recommendation: INCREASE social media budget by 20%
ROBUSTNESS CHECKS:
√ DiD estimate: 3.38% (consistent)
√ Leave-one-out: All estimates between 2.9% and 3.9%
√ Pre-period placebo: -0.21% (p = 0.84)
√ Compliance: 96% of Treatment DMAs within range
DATA QUALITY:
- Sales data completeness: 99.7%
- DMA assignment adherence: 100%
- Spillover detected: <2% in 4 Control DMAs
```

INTERPRET AND ACT

DECISION FRAMEWORK

Based on the results, follow your pre-committed decision rule:

RESULT	STATISTICAL SIGNIFICANCE	BUSINESS ACTION
LIFT ≥ TARGET	YES	SCALE SPEND PER PLAN
LIFT ≥ TARGET	NO	RERUN TEST WITH EXTENDED TIME OR SPEND
0 < LIFT < TARGET	YES	OPTIMIZE EFFICIENCY (CREATIVE, TARGETING)
0 < LIFT < TARGET	NO	CONSIDER REDUCING/REALLOCATING BUDGET
LIFT ≤ 0	ANY	PAUSE BUDGET AND INVESTIGATE FURTHER

COMMON INTERPRETATIONS

"We detected a 3.4% lift (p=0.029)"

- The channel works. Incremental revenue exceeds cost.
- Update MMM coefficients with this experimental prior
- Plan follow-up tests on sub-tactics (creative variants, audiences)

"Lift was 2.1% but not significant (p=0.18)"

- Effect may exist but test was underpowered
- Calculate: Would 2.1% lift be profitable? If yes, consider extending
- Review: Was compliance good? Any unusual events?

"No detectable lift (-0.5%, p=0.72)"

- Channel isn't moving the needle at current spend levels
- Before cutting: Verify test execution was clean
- Consider: Different creative? Different targeting? Or truly ineffective?

FEEDING BACK TO STRATEGY

UPDATE MARKETING MIX MODELS:

Use the experimental estimate as a Bayesian prior.

(For sample code, see Appendix A, Example 11: Update Marketing Mix Model with Experimental Prior)

Plan Next Tests:

Successful experiments often reveal new questions:

- If social works, which platform drives most lift?
- Does the effect vary by creative theme?
- Is there a frequency threshold?

Design follow-up experiments using the same rigorous framework.

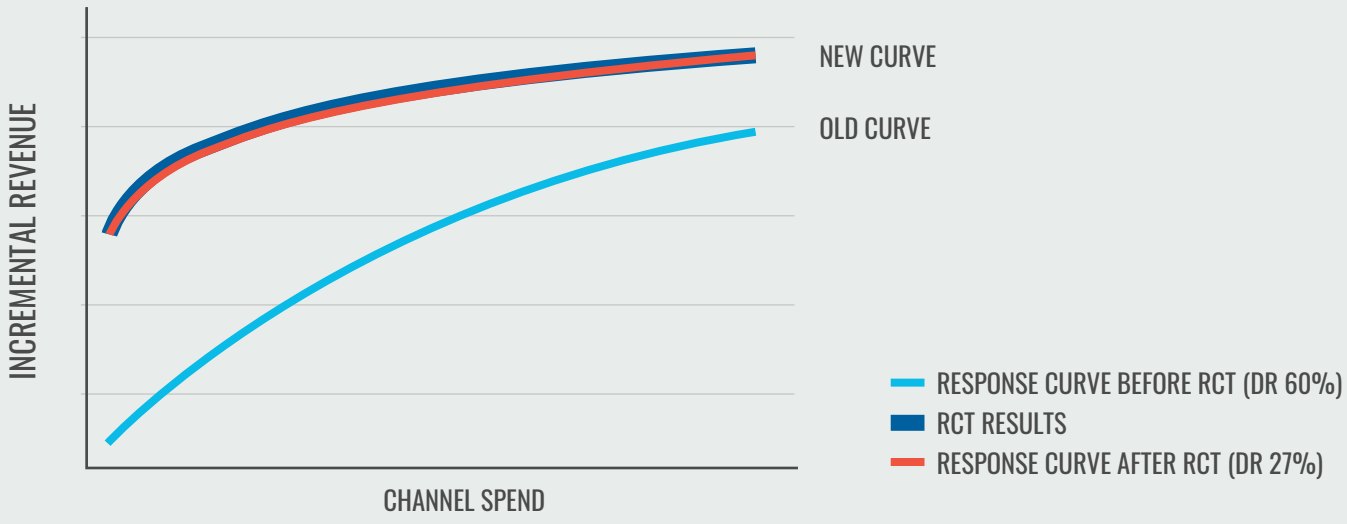
CLOSING THE LOOP

Document the full cycle:

1. Initial Hypothesis: “Social drives 3%+ lift”
2. Test Result: “Confirmed: 3.4% lift detected”
3. Business Decision: “Increased budget from $50MM to $60MM annually”
4. Actual Outcome: “6 months later: sustained iROAS of 2.8x on expanded budget”

This creates institutional knowledge and builds confidence in the experimentation program.

RECALIBRATING IMPACT RESPONSE CURVE IN MARKETING MIX MODEL



▲ **Illustration provided by MASS Analytics, MMM software provider.**
Compare the Calibrated Curve with the one previously used in the model. Use the calibrated Curve in lieu of the previous one. In this example we use DR (Diminishing Returns) 27% instead of 60%

COMMON PITFALLS AND HOW TO AVOID THEM

DESIGN PHASE PITFALLS

PITFALL: EXCLUDING "IMPORTANT" DMAS

Example:
"Let's exclude NYC and LA because they're too big to risk"

Why it's wrong:
Scope of sales effect estimate now only applies to smaller markets

Solution: Include all DMAs; let randomization handle heterogeneity.

PITFALL: PEEKING AT ASSIGNMENTS

Example:
"Treatment got more large DMAs, let's re-randomize"

Why it's wrong:
Not truly a randomized trail when treatment groups are cherry-picked[8]

Solution:
Design test ahead of time with stratification or re-randomization with pre-specified criteria

PITFALL: VAGUE SUCCESS CRITERIA

Example:
"We'll see if it works"

Why it's wrong:
Invites post-hoc rationalization

Solution:
Define specific lift threshold and decision rule upfront

EXECUTION PHASE PITFALLS

PITFALL: MID-FLIGHT ASSIGNMENT CHANGES

Example:
"Dallas is underdelivering, let's swap it to Control"

Why it's wrong:
Destroys randomization and causal inference

Solution:
Fix delivery issues with bid/budget changes only

PITFALL: CREATIVE VARIATIONS BY REGION

Example:
"Let's test the new creative in Treatment DMAs"

Why it's wrong:
Confounds media effect with creative effect

Solution:
Keep all non-media variables constant; Conduct creative testing or versioning as a separate experiment

PITFALL: REACTING TO EARLY RESULTS

Example:
"Week 2 looks great, let's double spending"

Why it's wrong:
Early results are noisy; changes compromise test

Solution:
Stick to the plan; save learnings for next test

[8] Optional stopping and the false discovery rate" (Hussey, 2018)

ANALYSIS PHASE PITFALLS

PITFALL: CHERRY-PICKING METRICS

Example:
"Revenue was flat but look at this segment!"

Why it's wrong:
Multiple testing without pre-specified design intent

Solution:
Pre-specify primary KPI; treat others as exploratory

PITFALL: IGNORING FAILED RANDOMIZATION

Example:
"Treatment had 20% higher pre-period sales but that's OK"

Why it's wrong:
Imbalance can masquerade as treatment effect

Solution:
Use normalized metrics or covariate adjustment

PITFALL: OVER-INTERPRETING MARGINAL RESULTS

Example:
"p=0.048, so it definitely works!"

Why it's wrong:
Threshold thinking ignores uncertainty

Solution:
Focus on confidence intervals and practical significance

ADVANTAGES OF LARGE-SCALE GEOGRAPHIC RANDOMIZED CONTROLLED EXPERIMENT

- **Scientifically sound:** It is a true Randomized Control Trial
- **Omnichannel:** It works for cable TV, CTV, digital video, programmatic, social, search, out-of-home, and more
- **Advertiser-friendly:** It is proven to work for advertisers, both large and small
- **Extensible:** Works for various KPIs including sales, foot traffic, TV tune-in and more
- **Independent:** Requires nothing but plan compliance from media firms, DSPs, agencies or other partners
- **Low Tech:** No special technology required (no clean rooms, user IDs, cookies, etc.)
- **Privacy assured:** No PII, only ZIP codes
- **Fraud proof:** Performance cannot be gamed: we defy anyone to hack it
- **Immediate results:** Final report in minutes of last KPI data availability
- **Simple:** Easily deployed in any media channel with DMA, ZIP code, or other geo targeting capabilities
- **Generalizable:** Projectable to national campaigns because it uses all DMAs in the country in the experiment
- **Transparent,** replicable, explainable

CONCLUSION

A well-designed geographic randomized controlled trial provides the gold standard for measuring advertising incrementality. By following this playbook, you can transform your organization's approach to marketing measurement and unlock significant competitive advantages.

BUILDING A CULTURE OF EXPERIMENTATION

The framework presented here – built on hundreds of real-world experiments – provides more than just a methodology. It represents a fundamental shift in how organizations should approach marketing ROI. While competitors continue relying on correlation-based attribution models and quasi-experimental methods that systematically overstate digital performance, companies that embrace geographic experimentation gain a decisive edge. In markets where share is zero-sum and growth comes at competitors' expense, understanding what truly drives incremental sales becomes your secret weapon, demonstrably changing perception of marketing from a cost center into a growth accelerator.

WHILE COMPETITORS CONTINUE RELYING ON CORRELATION-BASED ATTRIBUTION MODELS AND QUASI-EXPERIMENTAL METHODS THAT SYSTEMATICALLY OVERSTATE DIGITAL PERFORMANCE, COMPANIES THAT EMBRACE GEOGRAPHIC EXPERIMENTATION GAIN A DECISIVE EDGE.

STRATEGIC IMPLEMENTATION FOR MAXIMUM IMPACT

Start your experimentation journey with your largest spend categories, where even modest improvements in efficiency drive material business impact. A 5% improvement in a $100M channel delivers far more value than a 50% improvement in a $1M channel. But don't stop there – systematically work through your entire marketing mix:

1. **Channel Validation:** Test all major channels to establish true incremental contribution
2. **Risk Mitigation:** Never cut substantial spending without rigorous testing. Channels that appear ineffective in attribution models or with matched-market testing may be driving significant incremental sales
3. **Opportunity Discovery:** Investigate underinvested channels like audio, outdoor, and linear TV where attention may be high and CPMs low, potentially offering better returns than oversaturated digital channels
4. **Tactical Optimization:** After channel-level validation, test media partners, creative variants, audience segments, and messaging strategies
5. **Continuous Calibration:** Track whether mix changes deliver expected sales impact

FROM INSIGHTS TO INTELLIGENCE

As your organization builds a repository of experimental results – tens, then hundreds, eventually thousands of tests – you develop the industry's clearest picture of what drives incremental sales.

THIS EXPERIMENTAL BENCHMARK BECOMES YOUR COMPETITIVE MOAT. MAKE BOLD MOVES BACKED BY EXPERIMENTAL EVIDENCE WHILE COMPETITORS HESITATE.

Enabling:

- Predictive Modeling: Build AI models grounded in causal evidence rather than correlations (see "Predictive Incrementality by Experimentation"[9] for advanced applications)
- Strategic Confidence: Make bold moves backed by experimental evidence while competitors hesitate
- Compound Learning: Each experiment builds on previous insights, accelerating your pace of discovery

THE PATH FORWARD

Geographic RCTs offer simplicity in design but profound impact in results. By randomly assigning regions, delivering media, and measuring outcomes, you cut through the noise of modern marketing analytics. The careful attention to design details, statistical power, operational excellence, and analytical rigor pays dividends in decision quality.

This whitepaper demonstrates that while running experiments requires rigor and attention to detail, it's not as complicated or expensive as many marketers fear. This isn't rocket science, it's marketing science.

The real cost isn't in the resources required or the temporary sales suspension in control markets. The real cost is making million-dollar media decisions based on flawed measurement while billions in revenue opportunity hang in the balance.

Every day spent optimizing against inaccurate signals is a day your competitors might be using better evidence to steal your customers.

That said, executing geographic experiments well requires expertise, infrastructure, and sustained commitment. Not every organization needs to build this capability in-house. **Central Control stands ready to be your partner, bringing battle-tested methodology and platform capabilities to make geographic experimentation painless and optimal for your business.**

Remember: every failed experiment that prevents wasted spend is as valuable as a successful test that identifies a winning strategy. In the words often attributed to Thomas Edison, ""I have not failed. I've just found 10,000 ways that won't work."

Start today. Pick your largest channel. Design a test. Let real science, not assumptions or statistical mysticism, guide your next million-dollar decision. Soon, you'll be in the company of leading marketing experimentation experts like Netflix, Uber, Airbnb and other market leaders and wonder how you ever made decisions without this level of evidence.

[9] Predictive Incrementality by Experimentation (Gordon, et al., 2023)

APPENDICES

APPENDIX A: CODE EXAMPLES

EXAMPLE 1: SIMPLE RANDOMIZATION FOR GEOGRAPHIC RCT

Referenced in: Randomization Logic section

```
import pandas as pd
import numpy as np
# Set seed for reproducibility
np.random.seed(42)
# Read DMA list
dmas = pd.read_csv("dma_list.csv")
# Simple random assignment without replacement to avoid duplicate groupings
dmas["arm"] = np.random.choice(["Treatment", "Control"],
                               size=len(dmas),
                               replace=False)
# Check and print group size balance
group_counts = dmas["arm"].value_counts()
print("Group assignment counts:\n", group_counts)
# Save assignments
dmas.to_csv("geoRCT_assignments.csv", index=False)
```

EXAMPLE 2: POWER ANALYSIS DATA STRUCTURE

Referenced in: Historical Data Requirements for Simulation section

```
# Example data structure for power analysis
historical_data = pd.DataFrame({
'dma_code': [501, 501, 501, ...],
'week_ending': ['2023-01-07', '2023-01-14', ...],
'sales_volume': [125000, 132000, ...], # or
'transaction_count': [450, 475, ...],
'store_count': [12, 12, ...] # optional covariate
})
```

EXAMPLE 3: ESTIMATING AND APPLYING INTRACLASS CORRELATION COEFFICIENT (ICC)

Referenced in: Estimating and Applying Intraclass Correlation Coefficient (ICC) section

```
import pandas as pd
import numpy as np
from scipy.stats import ttest_ind
from statsmodels.regression.mixed_linear_model import MixedLM

# Simulate DMA-level sales data with a target ICC
def simulate_dma_sales(n_dmas=210, n_weeks=6, icc=0.15, mean_sales=1000,
                       sd_sales=300, seed=42):
    np.random.seed(seed)
    # Calculate between and within cluster variance
    var_total = sd_sales ** 2
    var_between = icc * var_total
    var_within = var_total - var_between
    # Generate random intercepts for each DMA
    dma_effects = np.random.normal(0, np.sqrt(var_between), size=n_dmas)
    # Simulate weekly sales per DMA
    data = []
    for dma in range(n_dmas):
        for week in range(n_weeks):
            y = mean_sales + dma_effects[dma] + np.random.normal(0,
np.sqrt(var_within))
            data.append({'dma': dma, 'week': week, 'sales': y})
    return pd.DataFrame(data)

# Estimate ICC using mixed-effects model (robust to imbalance)
def estimate_icc(df):
    model = MixedLM.from_formula('sales ~ 1', groups='dma', data=df)
    result = model.fit()
    var_between = result.cov_re.iloc[0, 0]   # Random intercept variance
    var_within = result.scale                # Residual (within-group)
variance
```

```
    icc = var_between / (var_between + var_within)
    return icc

# Compare power at different durations given ICC
def simulate_power(n_weeks_list=[4, 6], icc=0.15, lift=0.05, n_sim=100):
    results = []
    for weeks in n_weeks_list:
        significant = 0
        for _ in range(n_sim):
            df = simulate_dma_sales(n_weeks=weeks, icc=icc)
            treated = np.random.choice(df['dma'].unique(), size=105,
replace=False)
            df['group'] = df['dma'].apply(lambda x: 'T' if x in treated else
'C')

            # Apply lift to final week in treatment group
            df.loc[(df['group'] == 'T') & (df['week'] == weeks - 1),
'sales'] *= (1 + lift)

            # Difference-in-differences by DMA
            pre = df[df['week'] < weeks - 1].groupby('dma')['sales'].mean()
            post = df[df['week'] == weeks - 1].groupby('dma')['sales'].
mean()
            did = (post - pre).reset_index().merge(df[['dma', 'group']].
drop_duplicates(), on='dma')

            t, p = ttest_ind(did[did['group'] == 'T']['sales'],
                             did[did['group'] == 'C']['sales'],
                             equal_var=False)
            if p < 0.05:
                significant += 1
        power = significant / n_sim
        results.append({'weeks': weeks, 'power': power})
    return pd.DataFrame(results)
```

```
# Example usage
df = simulate_dma_sales(icc=0.18)
print(f"Estimated ICC: {estimate_icc(df):.3f}")

power_df = simulate_power(n_weeks_list=[4, 6, 8], icc=0.18)
print(power_df)
```

EXAMPLE 4: SQL TO PYTHON - EXTRACT RAW TRANSACTIONS

Referenced in: Data Preparation section

```
import pandas as pd
from sqlalchemy import create_engine

# Create database connection (replace with actual credentials)
engine = create_engine('your_connection_string')

# Define SQL query
query = """
SELECT
    customer_zip,
    transaction_date,
    sales_amount,           -- for volume
    1 AS trans_count        -- for counts
FROM transactions
WHERE transaction_date BETWEEN '2025-03-01' AND '2025-06-13'
  AND customer_zip IS NOT NULL
"""

# Execute and load into DataFrame
transactions = pd.read_sql(query, engine)
```

EXAMPLE 5: APPLY ZIP TO DMA MAPPING AND AGGREGATE

Referenced in: Data Preparation section

```
# Load ZIP to DMA mapping file
zip_dma_map = pd.read_csv('zip_dma_mapping.csv')
# Map ZIP codes to DMA codes
transactions = transactions.merge(
    zip_dma_map[['zip', 'dma_code']],
    left_on='customer_zip',
    right_on='zip',
    how='left'
)
# Convert transaction_date to datetime if needed
transactions['transaction_date'] = pd.to_datetime(transactions
['transaction_date'])
# Aggregate transactions to DMA-week level
weekly_data = transactions.groupby(
    ['dma_code', pd.Grouper(key='transaction_date', freq='W')]
).agg({
    'sales_amount': 'sum',   # Sum of sales for volume
    'trans_count': 'sum'     # Sum of transactions for counts
}).reset_index()
```

EXAMPLE 6: DATA QUALITY CHECKS

Referenced in: Data Quality Checks section

```
# Verify completeness
coverage = weekly_data.groupby('dma_code').size()
print(f"DMAs with full data: {(coverage == expected_weeks).sum()} / 210")
# Check for anomalies
import matplotlib.pyplot as plt
weekly_data.groupby('dma_code')['sales_amount'].plot(figsize=(12, 8))
plt.title('Sales by DMA Over Time')
plt.show()
# Test for skewness
from scipy import stats
```

```
skew = stats.skew(weekly_data['sales_amount'])
if abs(skew) > 1:
    print("Consider log or square-root transformation")
    weekly_data['log_sales'] = np.log(weekly_data['sales_amount'] + 1)

# If using sales volume (continuous), check distribution of normalized values.
# Normalize by population or baseline to remove DMA size effects
# Example: sales per capita or index to baseline period
weekly_data['sales_per_capita'] = weekly_data['sales_amount'] / weekly_data['population']
# Or normalize to baseline period (e.g., pre-treatment mean)
baseline_means = weekly_data[weekly_data['week'] < treatment_start].groupby('dma')['sales_amount'].mean()
weekly_data['sales_index'] = weekly_data.apply(lambda x: x['sales_amount'] / baseline_means[x['dma']], axis=1)
# Test normalized values for approximate normality
skew = stats.skew(weekly_data['sales_per_capita'])
if abs(skew) > 1:
    print(f"Normalized sales skewness: {skew:.2f}")
    print("Consider additional transformations if needed for model assumptions")
```

EXAMPLE 7: PRIMARY ANALYSIS - T-TEST ON NORMALIZED VALUES

Referenced in: Primary Analysis section

```
import pandas as pd
import numpy as np
from scipy import stats
import statsmodels.api as sm
from statsmodels.stats.anova import anova_lm
from statsmodels.stats.sandwich_covariance import cov_hc1
# Load DMA-level pre/post sales and assignment data
df = pd.read_csv("geo_rct_results.csv")
# Step 1: Calculate pre-period metrics (average + trend)
pre_period = df[df['period'] == 'pre'].groupby('dma_code').agg({
```

```
    'weekly_sales': [
        'mean',
        lambda x: np.polyfit(range(len(x)), x, 1)[0] # slope of sales trend
    ]
}).reset_index()
pre_period.columns = ['dma_code', 'pre_avg', 'pre_trend']
# Step 2: Calculate average test-period sales
test_period = df[df['period'] == 'test'].groupby('dma_code').agg({
    'weekly_sales': 'mean'
}).reset_index()
test_period.columns = ['dma_code', 'test_avg']
# Step 3: Merge pre, test, and assignment groups
analysis_df = pre_period.merge(test_period, on='dma_code')
analysis_df = analysis_df.merge(
    df[['dma_code', 'assignment']].drop_duplicates(),
    on='dma_code'
)
# Step 4: Estimate expected sales using pre-period trend
# Assumes linear growth and 5 weeks of test period
analysis_df['expected_sales'] = analysis_df['pre_avg'] * (1 + analysis_
df['pre_trend'] * 5)
# Step 5: Calculate normalized lift index
analysis_df['lift_index'] = (analysis_df['test_avg'] / analysis_
df['expected_sales']) - 1
# Step 6: Run OLS regression with treatment group as predictor
X = sm.add_constant(pd.get_dummies(analysis_df['assignment'], drop_
first=True))  # e.g., Control=0, Treatment=1
y = analysis_df['lift_index']
model = sm.OLS(y, X).fit()
print(model.summary())
# Step 7: Report robust (HC1) standard errors
robust_cov = cov_hc1(model)
robust_se = np.sqrt(np.diag(robust_cov))
print(f"Robust standard error for treatment effect: {robust_se[1]:.4f}")
```

EXAMPLE 8: DIFFERENCE-IN-DIFFERENCES CROSS-CHECK

Referenced in: Robustness Checks section

```
import pandas as pd
import statsmodels.formula.api as smf
# Copy original data and create post-treatment indicator
did_df = df.copy()
did_df['post'] = (did_df['period'] == 'test').astype(int)
# Run Difference-in-Differences model with fixed effects and clustered SEs
did_model = smf.ols(
    formula='weekly_sales ~ assignment * post + C(dma_code)',  # includes
DMA fixed effects
    data=did_df
).fit(
    cov_type='cluster',
    cov_kwds={'groups': did_df['dma_code']}  # cluster SEs at DMA level
)
# Output the DiD interaction coefficient
interaction_term = 'assignment[T.Treatment]:post'
if interaction_term in did_model.params:
    print(f"DiD estimate: {did_model.params[interaction_term]:.4f}")
else:
    print("Interaction term not found in model output. Check data
encoding.")
```

EXAMPLE 9: LEAVE-ONE-OUT ANALYSIS

Referenced in: Robustness Checks section

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
# Run leave-one-out regression by dropping one DMA at a time
loo_results = []
for excluded_dma in analysis_df['dma_code'].unique():
    temp_df = analysis_df[analysis_df['dma_code'] != excluded_dma]
```

```
    # Regress lift_index ~ assignment (Treatment vs Control)
    model_loo = sm.OLS(
        temp_df['lift_index'],
        sm.add_constant(pd.get_dummies(temp_df['assignment'], drop_
first=True))
    ).fit()
    loo_results.append({
        'excluded_dma': excluded_dma,
        'estimate': model_loo.params[1],    # treatment coefficient
        'pvalue': model_loo.pvalues[1]      # p-value for treatment effect
    })
# Create dataframe of LOO estimates
loo_df = pd.DataFrame(loo_results)
# Plot LOO estimates to identify influential DMAs
plt.figure(figsize=(10, 6))
plt.scatter(loo_df['excluded_dma'], loo_df['estimate'], alpha=0.7)
plt.axhline(y=0.0342, color='red', linestyle='--', label='Full-sample
estimate')
plt.xlabel('Excluded DMA')
plt.ylabel('Treatment Effect Estimate')
plt.title('Leave-One-Out Sensitivity Analysis')
plt.legend()
plt.tight_layout()
plt.show()
```

EXAMPLE 10: PRE-PERIOD BALANCE CHECK

Referenced in: Robustness Checks section

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
# Create fake pre/test periods using pre-period weeks
placebo_df = df[df['period'] == 'pre'].copy()
placebo_df['fake_period'] = np.where(
    placebo_df['week_number'] >= 5, 'fake_test', 'fake_pre'
```

```
)
# Group by DMA and fake period to get average sales
placebo_summary = placebo_df.groupby(['dma_code', 'fake_period'])['weekly_
sales'].mean().reset_index()
# Pivot to wide format
placebo_pivot = placebo_summary.pivot(
    index='dma_code',
    columns='fake_period',
    values='weekly_sales'
).reset_index()
# Add treatment/control assignment
placebo_pivot = placebo_pivot.merge(
    df[['dma_code', 'assignment']].drop_duplicates(),
    on='dma_code'
)
# Compute placebo lift (should be ~0 if pre-periods are balanced)
placebo_pivot['fake_lift'] = (placebo_pivot['fake_test'] / placebo_
pivot['fake_pre']) - 1
# Run placebo regression
placebo_model = sm.OLS(
    placebo_pivot['fake_lift'],
    sm.add_constant(pd.get_dummies(placebo_pivot['assignment'], drop_
first=True))
).fit()
# Report placebo results
print(f"Placebo effect: {placebo_model.params[1]:.4f}")
print(f"Placebo p-value: {placebo_model.pvalues[1]:.4f}")
```

EXAMPLE 11: UPDATE MARKETING MIX MODEL WITH EXPERIMENTAL PRIOR

Referenced in: Feeding Back to Strategy section

```
# Example: Updating MMM with experimental prior
# Prior from experiment: 3.4% ± 1.5%
prior_mean = 0.034
prior_sd = 0.015
```

```
# Combine with MMM posterior using inverse variance weighting
mmm_estimate = 0.028
mmm_se = 0.022
pooled_estimate = (
    (prior_mean / prior_sd**2 + mmm_estimate / mmm_se**2) /
    (1 / prior_sd**2 + 1 / mmm_se**2)
)
pooled_se = np.sqrt(1 / (1 / prior_sd**2 + 1 / mmm_se**2))
print(f"Updated coefficient: {pooled_estimate:.3f}")
print(f"Updated standard error: {pooled_se:.3f}")
```

EXAMPLE 12: STRATIFIED RANDOMIZATION

Referenced in: Alternative Randomization Approaches section

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
def stratified_randomization(dmas_df, strat_vars, n_strata=4, seed=42):
    """
    Stratified randomization for Geo RCT
    """
    # Standardize stratification variables
    scaler = StandardScaler()
    X = scaler.fit_transform(dmas_df[strat_vars])
    # Create strata using k-means clustering
    kmeans = KMeans(n_clusters=n_strata, random_state=seed)
    dmas_df['stratum'] = kmeans.fit_predict(X)
    # Randomize within strata
    np.random.seed(seed)
    dmas_df['assignment'] = 'Control'
    for stratum in range(n_strata):
        stratum_indices = dmas_df[dmas_df['stratum'] == stratum].index
        n_treat = len(stratum_indices) // 2
```

```
        treat_dmas = np.random.choice(stratum_indices, n_treat,
replace=False)
        dmas_df.loc[treat_dmas, 'assignment'] = 'Treatment'
        # Handle the leftover DMA if stratum size is odd
        if len(stratum_indices) % 2 == 0:
            remaining = list(set(stratum_indices) - set(treat_dmas))
            assign_to = np.random.choice(['Treatment', 'Control'])
            dmas_df.loc[np.random.choice(remaining, 1), 'assignment'] =
assign_to
    return dmas_df
# Example usage
dmas = pd.read_csv('dma_data.csv')
strat_vars = ['pre_period_sales', 'population', 'median_income']
assignments = stratified_randomization(dmas, strat_vars)
```

EXAMPLE 13: POWER SIMULATION FRAMEWORK

Referenced in: Simulation-Based Power Estimation section

```
import pandas as pd
import numpy as np
from joblib import Parallel, delayed
from scipy.stats import ttest_ind
def run_power_simulation(historical_data, effect_sizes=[0.02, 0.03, 0.05],
                         test_weeks=[3, 4, 5, 6, 8], n_sims=1000,
alpha=0.05):
    """
    Run power simulation for geographic RCT
    """
    results = []
    for effect in effect_sizes:
        for weeks in test_weeks:
            # Run simulations in parallel
            sim_results = Parallel(n_jobs=-1)(
                delayed(single_simulation)(
```

```
                    historical_data, effect, weeks, alpha
                ) for _ in range(n_sims)
            )
            power = np.mean(sim_results)
            results.append({
                'effect_size': effect,
                'test_weeks': weeks,
                'power': power
            })
    return pd.DataFrame(results)
def single_simulation(historical_data, effect, weeks, alpha):
    """
    Single simulation iteration
    """
    # Sample random test window (ensure space for 8-week pre-period)
    start_week = np.random.randint(8, len(historical_data) - weeks)
    # Extract pre and test periods
    pre_data = historical_data.iloc[start_week - 8:start_week].copy()
    test_data = historical_data.iloc[start_week:start_week + weeks].copy()
    # Random assignment of DMAs to treatment and control
    dmas = historical_data['dma_code'].unique()
    treatment_dmas = np.random.choice(dmas, len(dmas) // 2, replace=False)
    pre_data['group'] = pre_data['dma_code'].apply(lambda x: 'T' if x in
treatment_dmas else 'C')
    test_data['group'] = test_data['dma_code'].apply(lambda x: 'T' if x in
treatment_dmas else 'C')
    # Aggregate weekly sales by DMA
    pre_avg = pre_data.groupby('dma_code')['sales_volume'].mean().reset_
index(name='pre_avg')
    test_avg = test_data.groupby('dma_code')['sales_volume'].mean().reset_
index(name='test_avg')
    # Apply simulated lift to treatment group
    test_avg['group'] = test_avg['dma_code'].apply(lambda x: 'T' if x in
treatment_dmas else 'C')
    test_avg.loc[test_avg['group'] == 'T', 'test_avg'] *= (1 + effect)
```

```
    # Merge for lift calculation
    merged = pre_avg.merge(test_avg[['dma_code', 'test_avg', 'group']],
on='dma_code')
    # Compute lift index
    merged['lift'] = (merged['test_avg'] / merged['pre_avg']) - 1
    # Run t-test on lift between groups
    t_vals = merged[merged['group'] == 'T']['lift']
    c_vals = merged[merged['group'] == 'C']['lift']
    t_stat, p_val = ttest_ind(t_vals, c_vals, equal_var=False)
    return p_val < alpha
# Example usage
# historical_data = pd.read_csv('historical_sales_data.csv')
# power_results = run_power_simulation(historical_data)
# print(power_results)
```

APPENDIX B: GLOSSARY

ANCOVA: Analysis of Covariance; regression model that includes treatment indicator and continuous covariates

Cluster-RCT: Cluster Randomized Controlled Trial; experimental design where groups (clusters) rather than individuals are randomly assigned to treatment conditions

DMA: Designated Market Area; geographic regions defined by Nielsen for television viewership measurement, commonly used as experimental units in geo tests

ICC: Intraclass Correlation Coefficient; measures the similarity of observations within the same cluster relative to observations between clusters

iROAS: Incremental Return on Ad Spend; the causal revenue impact per dollar spent, measured through experimentation rather than correlation

ITT: Intent-to-Treat; analysis based on initial treatment assignment regardless of actual exposure, preserving the benefits of randomization

MDE: Minimum Detectable Effect; the smallest true effect size that an experiment has adequate power to detect as statistically significant

MECE: Mutually Exclusive and Collectively Exhaustive; property where categories don't overlap and cover all possibilities

MMM: Marketing Mix Model; statistical model that decomposes sales into contributions from various marketing channels and external factors

SMD: Standardized Mean Difference; difference in means divided by pooled standard deviation, used to assess balance between groups

SUTVA: Stable Unit Treatment Value Assumption; assumption that treatment of one unit doesn't affect outcomes of other units (no spillover)

APPENDIX C: PRE-LAUNCH CHECKLIST

Business Alignment

- ☐ Business objective clearly stated
- ☐ Decision rule documented (if X then Y)
- ☐ Budget approved and ring-fenced
- ☐ Stakeholders signed off on design

Technical Setup

- ☐ Historical data validated (2+ years)
- ☐ Power analysis completed
- ☐ Randomization code peer-reviewed
- ☐ DMA assignments generated and locked
- ☐ Version control established

Media Readiness

- ☐ Platform targeting lists created
- ☐ Creative assets approved and identical
- ☐ Trafficking instructions documented
- ☐ Compliance monitoring dashboard live
- ☐ Pacing alerts configured

Data Infrastructure

- ☐ KPI data pipeline tested
- ☐ ZIP→DMA mapping current
- ☐ Latency documented and acceptable
- ☐ Analysis code templates ready
- ☐ Results template prepared

APPENDIX D: POST-TEST CHECKLIST

Data Quality

- ☐ Sales data completeness >99%
- ☐ No DMA assignment violations
- ☐ Spillover quantified and <5%
- ☐ Outliers investigated
- ☐ Time series plots reviewed

Analysis Completeness

- ☐ Primary t-test completed
- ☐ DiD cross-check performed
- ☐ Leave-one-out sensitivity done
- ☐ Pre-period placebo tested
- ☐ Confidence intervals calculated

Documentation

- ☐ Results summary drafted
- ☐ Technical appendix complete
- ☐ Visualizations created
- ☐ Recommendations clear
- ☐ Lessons learned captured

Action Items

- ☐ Decision rule executed
- ☐ Budget changes implemented
- ☐ MMM coefficients updated
- ☐ Next test planned
- ☐ Results socialized

ABOUT CENTRAL CONTROL

Central Control helps advertisers and their partners measure the sales impact of advertising using rigorous, science-based methods that are independent, transparent, unbiased, replicable, and explainable. We offer a platform, consulting, and training to support high-quality advertising experiments that accurately measure incremental return on ad spend (iROAS).

Founded in 2020, the company works with brands, media companies, agencies, and others to build a clearer understanding of advertising's true effect on sales and other business outcomes. We specialize in helping organizations reframe how they assess advertising effectiveness for competitive advantage, offering executive advising, measurement retooling, and calibration of marketing mix models.

The Central Control team includes ad industry veterans from Google, DoubleClick, Microsoft, Amazon, and Adobe, with deep expertise in experimental design, media measurement, data science, marketing analytics, econometrics, and applied research.

Our team has supported 500+ experiments for Fortune 500 advertisers, optimizing billions of dollars in business impact.

CONTACT:

Email: info@centralcontrol.com

Web: www.centralcontrol.com

LinkedIn: /company/centralcontrol

ABOUT THE AUTHOR:

Rick Bruner is CEO and founder of Central Control. He has spent 25+ years at the intersection of advertising and technology, previously running research and product departments at DoubleClick, Google, Viacom, Marketing Evolution, Viant and Guideline. He is Vice Chair for the marketing sciences industry body I-COM and founder and moderator of the influential Research Wonks online forum.

ACKNOWLEDGEMENTS:

Various contributors provided valuable input to the production of this paper, namely these individuals: John Chandler, PhD, Head of Data Science at Central Control, Managing Partner of Data Insights LLC and Clinical Professor of Marketing at the University of Montana, for help in designing many of these experimental techniques and technical review of the paper; Kumi Harischandra, research scientist, for technical review of the paper; Campbell Foster, Chief Commercial Office of Central Control, for editing, and Ben Munday, Creative Director of Munday Design, for graphic design.