Learning from Many Experiments: A Hierarchical Bayesian Framework for Decomposing Marketing Treatment Heterogeneity

January 25, 2026

Abstract

Firms increasingly rely on A/B testing to evaluate marketing strategies, yet most experiments are analyzed in isolation, limiting insight into why effectiveness varies and how repeated exposure shapes outcomes. We develop a hierarchical Bayesian framework that jointly analyzes randomized marketing interventions to decompose treatment effect heterogeneity into three components: customer responsiveness, campaign design, and contextual timing. The model is scalable, incorporates both observed and unobserved variation, and links estimation directly to policy evaluation with full uncertainty quantification. Using data from large-scale field experiments involving nearly half a million customers, we document three main findings. First, unobserved customer-level heterogeneity accounts for the majority of variation in treatment effects, explaining roughly two-thirds of the total dispersion, far exceeding the contribution of campaign design or timing. Second, we find strong evidence of intervention fatigue: responsiveness declines with repeated exposure and recovers only slowly. Third, in a held-out policy evaluation, a model-based targeting strategy improves revenue on average while substantially reducing the share of customers targeted, with gains reaching over 10% in settings where experimental variation is most informative. Approximately 80% of these gains are driven by personalization based on latent customer responsiveness rather than observable characteristics. Together, these results demonstrate how repeated experimentation can be leveraged to explain variation in marketing effectiveness and to support scalable targeting and personalization.

Keywords: Marketing interventions, heterogeneity, field experiments, Bayesian modeling, treatment effect decomposition, targeting.

1. Introduction

Understanding the impact of marketing interventions is a central challenge in marketing. In an era characterized by ubiquitous experimentation and increasing opportunities for personalization, firms frequently deploy A/B tests to evaluate and optimize their marketing strategies. Despite the widespread adoption of these experiments, the effectiveness of marketing actions remains highly variable across campaigns, consumers, and over time (e.g., Lodish et al. 1995; Gordon et al. 2019; Ellickson, Kar, and Reeder III 2023). For marketing managers, understanding the sources of this variation is critical for making informed decisions regarding resource allocation, campaign design, and the development of effective personalization strategies.

Notwithstanding the rise of large-scale experimentation in domains such as e-commerce, telecommunications, web-design and digital marketing, current practice tends to analyze each campaign in isolation (e.g., Lodish et al. 1995; Kohavi and Longbotham 2023; Ellickson, Kar, and Reeder III 2023). While this approach provides an internally valid assessment of causal effects, it does not leverage the possible insights that can be gained and actions that can be taken by jointly analyzing information across campaigns. A cross-campaign perspective allows researchers and managers to uncover systematic patterns that are difficult or impossible to discern when experiments are evaluated independently. Specifically, it can shed light on several substantive and strategic questions.

First, what are the dominant sources of variation in marketing effectiveness? Are differences in outcomes primarily driven by campaign design, such as the promotion offered or targeting schema, by the timing of the campaign, or the preferences of the customers being targeted? Second, to what extent is this variation systematic and therefore predictable? For instance, are some customers inherently more responsive to marketing, perhaps due to unobserved characteristics or strategic behavior that enable them to extract more value from firm-initiated interventions? Third, what are the dynamic consequences of repeated exposure to marketing interventions over time — specifically,

how responsiveness varies with customers' prior exposure histories (Mela, Jedidi, and Bowman 1998; Kopalle, Mela, and Marsh 1999)? Does the increasing contact frequency, facilitated by ongoing experimentation, diminish customer responsiveness and, in turn, reduce the long-run efficacy of marketing efforts?

This paper addresses these questions by leveraging data from repeated randomized interventions and estimating a hierarchical Bayesian model that jointly analyzes the heterogeneous treatment effects within and across campaigns and decomposes them across three distinct dimensions: the design of the campaign, the contextual timing of its delivery, and the intrinsic responsiveness of individual customers. The proposed framework allows us to jointly estimate these components, accommodating both observed and unobserved sources of heterogeneity. The Bayesian approach is particularly well suited to this context, as it naturally incorporates uncertainty and enables the recovery of latent structure from sparse or unbalanced distributed exposure histories. We apply this framework to a comprehensive dataset from a large telecommunications provider, including more than a hundred randomized marketing interventions over a two-year period. The richness and scale of these data permit a fine-grained investigation of how responsiveness varies across individuals, campaigns, and time.

The empirical results yield several key findings. First, we document that the majority of treatment effect variation across campaigns is attributable to customer-level heterogeneity. This heterogeneity is largely unobserved, persisting even after conditioning on a broad set of behavioral covariates. Second, while campaign design and contextual timing contribute meaningfully to variation in effectiveness, their relative importance is substantially smaller. Third, we provide strong evidence of behavioral saturation (Mela, Jedidi, and Bowman 1998; Kopalle, Mela, and Marsh 1999; Sahni 2015): customer responsiveness declines with repeated exposure to interventions, and recovery over time is modest at best. This pattern highlights the risks associated with frequent experimentation and targeting and underscores the need for strategic pacing of marketing contact.

Importantly, we also demonstrate that these insights have material implications for targeting. A policy that leverages our model to prioritize customers with positive predicted treatment effects improves performance relative to both the firm's existing policy and a benchmark that assigns treatment at random. While average gains are modest when aggregated across all campaigns, they vary substantially across settings. In particular, wIn particular, we find larger targeting gains for campaigns that encouraged a behavioral response from the customer, like re-engagement or cross-sell, and for interventions that maximize the exploitation of insights through a balanced randomization, leading to over 10% gains for certain campaigns.e gains reveals that approximately 83% of the improvement is driven by unobserved customer-level responsiveness, rather than by observable campaign characteristics or timing effects. These results underscore the value of integrating information across experiments to uncover latent sources of responsiveness that are inaccessible to standard per-campaign analyses.

In sum, our findings highlight the value of leveraging information across repeated experimentation both for a better understanding of *what* makes a campaign work, but also to uncover *how* such insights can be translated into actionable personalization strategies. By modeling repeated experimentation as a unified data-generating process rather than a collection of isolated tests, our framework identifies the key drivers of effectiveness, quantifies their relative importance, and informs scalable targeting policies that account for both observed and unobserved heterogeneity. Together, these insights position repeated experimentation as a foundation for more systematic learning and more effective personalization in marketing practice.

Collectively, this research makes three primary contributions. Substantively, we advance the literature on marketing effectiveness by leveraging multiple experiments to decompose treatment effect heterogeneity into interpretable components, allowing us to quantify the relative contributions of customer-level variation, campaign design, and contextual timing. While prior work has widely acknowledged heterogeneity in response to

marketing interventions (e.g., Zantedeschi, Feit, and Bradlow 2017; Gordon et al. 2019; Ellickson, Kar, and Reeder III 2023; Hitsch, Misra, and Zhang 2024), systematic measurement of its *sources* and their *relative importance* has been rare. Existing studies and meta-analyses primarily partition variation by campaign/channel or context, for example, decomposing promotional "bumps," explaining cross-study elasticity differences, or comparing mix elements and multichannel effects (van Heerde, Leeflang, and Wittink 2004; Bijmolt, van Heerde, and Pieters 2005; Ataman, van Heerde, and Mela 2010), not considering *customer-level* heterogeneity. Our analysis fills this gap and shows that the dominant source of variation is unobserved customer-level heterogeneity, alongside behavioral dynamics from repeated exposure, thereby informing theories of customer heterogeneity, intervention fatigue, and the strategic pacing of marketing campaigns.

From a managerial perspective, we demonstrate how firms can benefit from aggregating insights across multiple experiments, rather than evaluating each A/B test in isolation. Our approach estimates individual-level responsiveness to marketing interventions and enables more effective data-driven targeting strategies (Rossi, McCulloch, and Allenby 1996; Ascarza 2018; Ellickson, Kar, and Reeder III 2023; Hitsch, Misra, and Zhang 2024). We show that meaningful performance gains in incremental revenue can be achieved through cross-experimental analysis.

Methodologically, we propose a scalable Bayesian framework that not only estimates treatment effect heterogeneity but also links these estimates directly to policy evaluation. To our knowledge, this is among the first approaches to explicitly combine Bayesian inference with large-scale experimental analysis for marketing applications. By looking across many experiments, our approach combines the causal-inference advantages of randomized experimentation with the longitudinal richness of repeated observational data. Bayesian methodologies are particularly effective in such settings, as they can pool information across related observations to capture unobserved sources of heterogeneity, dynamic structures, and the multiple drivers underlying observed outcomes (Rossi, Mc-

Culloch, and Allenby 1996). Hence, our approach leverages the strengths of Bayesian estimation for repeated experimentation, using hierarchical structure to integrate information across interventions while preserving the internal validity of each experiment. By combining these elements, our framework extends existing off-policy evaluation methods by incorporating full posterior uncertainty, allowing decision-makers to quantify both the expected value and the associated risk of alternative targeting policies. As a result, our approach connects model predictions to actionable business outcomes. Whereas prior research typically relies on point estimates or deterministic model-based evaluations, our framework produces posterior distributions of policy outcomes based on holdout data, enabling more robust and risk-aware decision-making.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 outlines the empirical context and data. Section 4 introduces the hierarchical Bayesian framework, including model specification, estimation procedure, and validation strategy. Section 5 reports the main findings and evaluates their implications for targeting. Section 7 decomposes treatment effect heterogeneity into customer, campaign design, and contextual factors. Section 8 concludes with implications and limitations.

2. Literature

This research builds on and extends several streams of work in marketing, including studies of marketing effectiveness decomposition, long-term effects of promotion sensitivity, heterogeneous treatment effects, personalization, and learning from past experiments.

Variation and Dynamics of Promotion Sensitivity A central theme in marketing research is understanding the effectiveness of promotional interventions and the variation in their impact across individuals and contexts. Prior work documents how marketing actions influence outcomes such as brand switching, purchase timing, and consumption incidence (Gupta 1988; Bell, Chiang, and Padmanabhan 1999; Chan, Narasimhan, and

Zhang 2008), but has paid less attention to decomposing the *sources* of that variation in a causal-experimental setting. We shift the focus from decomposing behavioral outcomes to decomposing *treatment effects*, quantifying how much of the variation in promotional effectiveness arises from differences across consumers, intervention designs, and contextual timing. This distinction is critical: it moves beyond describing what happened to explaining *what* features of the campaign, timing, or audience make seemingly similar interventions yield divergent results.

A large body of literature also shows that repeated exposure generates dynamic responses (e.g., stockpiling, reduced effectiveness, delayed purchases) (e.g., Mela, Gupta, and Lehmann 1997; Mela, Jedidi, and Bowman 1998; Kopalle, Mela, and Marsh 1999), with the magnitude depending on timing and frequency (Gupta 1988). While these dynamics are often studied in secondary data, we extend this work to the realm of repeated field experiments, showing that responsiveness systematically deteriorates with repeated interventions, even when designs differ, highlighting *intervention saturation*. Our findings echo nonlinear exposure effects in advertising (Sahni 2016) and aligns with recent evidence of behavioral fatigue in sequential experiments (Shchetkina and Berman 2024).

Heterogeneity in Treatment Effects and Unobserved Responsiveness The personalization literature increasingly estimates heterogeneous treatment effects (HTEs) to guide targeting (e.g., Ascarza 2018; Imai and Li 2021; Yoganarasimhan, Barzegary, and Pani 2023). Recent advances use ML to estimate CATEs and evaluate policies based on predicted responsiveness (Ellickson, Kar, and Reeder III 2023; Hitsch, Misra, and Zhang 2024; Huang and Ascarza 2024). However, most studies analyze one experiment at a time, thereby emphasizing observed heterogeneity and obscuring how much variation stems from *customers* versus *campaign design* or *delivery context*. Our approach complements this work by jointly analyzing many experiments and explicitly decomposing HTEs into customer-, campaign-, and context-level components. Crucially, the multi-experiment,

hierarchical Bayesian design allows us to recover *unobserved* individual responsiveness (beyond observed covariates) and to use it for targeting, addressing calls to open the "black box" of omitted moderators (Krefeld-Schwalb, Sugerman, and Johnson 2024). This emphasis on unobserved, persistent responsiveness naturally motivates a hierarchical Bayesian framework, which provides the structure necessary to capture such latent heterogeneity and to propagate uncertainty through to decision-making; as we discuss in greater detail below.

Leveraging Multiple Experiments for Learning and Personalization While most field-experimental research analyzes each experiment in isolation, several recent studies have begun to explore how information can be shared across experiments to enhance inference and targeting performance. Zantedeschi, Feit, and Bradlow (2017) develop a hierarchical Bayesian model of multichannel advertising response that pools data across repeated holdout campaigns to estimate channel-specific carryover and cross-channel synergies. While their focus is on understanding advertising dynamics across channels, our approach similarly leverages repeated randomized interventions but has a different objective: rather than modeling channel carryover, we use cross-experimental pooling to *decompose* treatment-effect heterogeneity into three interpretable components—customer, campaign, and temporal context—and to link these components directly to policy evaluation.

Ellickson et al. (2025) extend causal prediction to *unstructured treatments* such as email subject lines. Their framework encodes textual content via contextual embeddings and combines these embeddings with doubly robust causal scores to predict the performance of novel messages, integrating generative AI to create and evaluate new content. Our work addresses a distinct problem. We study *structured marketing interventions* repeatedly randomized over time and use a hierarchical Bayesian decomposition to quantify *where*

and why effectiveness varies, across customers, designs, and timing, rather than predicting outcomes for new, untested creative content.

Huang, Ascarza, and Israeli (2024) and Ibragimov et al. (2025) also leverage information across experiments, but do so through transfer-learning approaches. Huang, Ascarza, and Israeli (2024) propose a two-stage causal machine-learning framework that synthesizes past experiments using doubly robust scores and deep representation learning, enabling prediction of effects for new interventions. Ibragimov et al. (2025) develop a Bayesian probabilistic matrix factorization method that transfers knowledge across campaigns at the segment level, explicitly accounting for uncertainty in pre-estimated segment-level treatment effects. Our framework differs from these approaches in both *unit of analysis* and *goal*. We estimate individual-level treatment effects directly from raw randomized data across many interventions, rather than relying on segment-level summaries or observed design features. Instead of predicting or transferring outcomes to new interventions, we provide a *decomposition-based explanation* of what drives variation, identifying whether transferable structure arises primarily from customer responsiveness, campaign design, or contextual timing. This decomposition yields interpretable diagnostics on where learning from prior experiments generalizes and where it does not, thereby informing both targeting and pacing decisions.

In this sense, our study contributes a *complementary and interpretable perspective* on cross-experimental learning. By isolating the relative contributions of different sources of heterogeneity, we clarify what lessons from past experiments are stable and actionable. This aligns with the broader call for using experimentation as a foundation for personalization under uncertainty (Athey, Wager, and Syrgkanis 2021) and complements counterfactual-modeling approaches that emphasize adaptive targeting and design (Dell'Acqua, Cohen, and Zhou 2021).

Bayesian Perspective and Policy Evaluation Hierarchical Bayesian models have long been used in marketing to capture individual heterogeneity and improve estimation efficiency (e.g., Rossi, McCulloch, and Allenby 1996; Zantedeschi, Feit, and Bradlow 2017). We build on this tradition by extending Bayesian estimation to the context of large-scale experimental data and linking it explicitly to policy evaluation. Our framework propagates *posterior uncertainty* from the estimation of heterogeneous treatment effects to the evaluation of targeting policies, allowing decision-makers to quantify both expected performance and its associated uncertainty. For validation and policy-value assessment, we employ rank-weighted estimands such as the Targeting Operating Characteristic (TOC) and its area under the curve (AUTOC; Yadlowsky et al. 2025), which measure how effectively the model ranks customers by expected incremental impact. Whereas most off-policy evaluations rely on point estimates or plug-in frequentist metrics, our Bayesian formulation produces a posterior distribution of policy outcomes, supporting more robust and risk-aware decision-making.

In sum, our work differs from existing multi-experiment and transfer-learning approaches by (i) estimating *individual-level causal effects* across many randomized interventions and *decomposing* their variation into customer, campaign, and temporal components, including unobserved responsiveness, and (ii) connecting this decomposition directly to *posterior, uncertainty-aware policy evaluation*. This combination provides a scalable and interpretable foundation for learning from repeated experimentation and for designing targeting and pacing policies that balance value creation with confidence in expected outcomes.

3. Empirical Setting and Data

Empirical setting. We use data from a large telecommunications provider in the Middle East that regularly sends promotional text-based SMS offers to its customers. We fo-

cus on its prepaid mobile segment, which constitutes the majority of the firm's customer base. A key feature of this empirical setting is the firm's systematic use of randomized experimentation: virtually all marketing campaigns include a randomized holdout group, providing a unique opportunity to examine the impact of repeated interventions at scale.

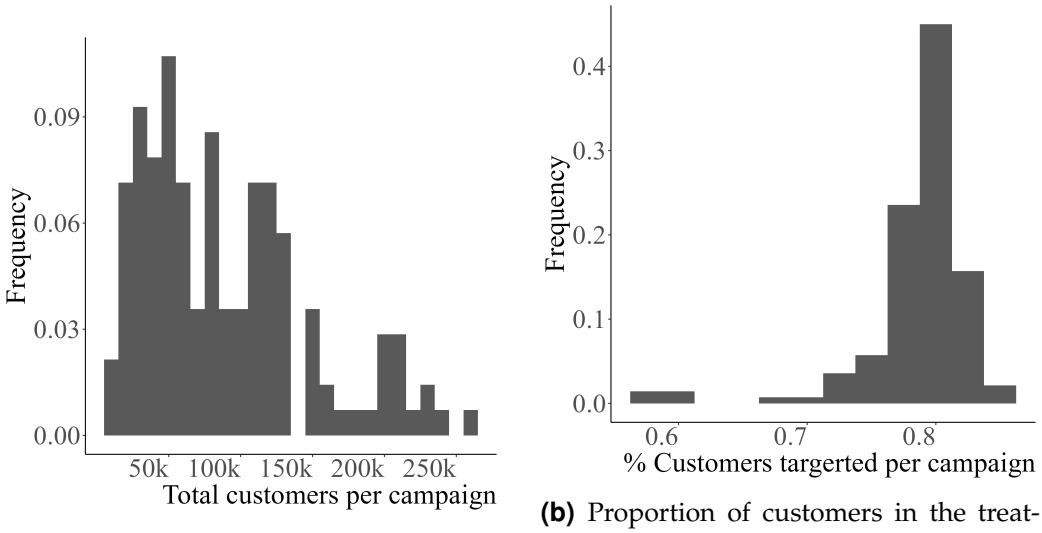
Data overview. The dataset spans approximately two years (96 weeks) and includes about 580,000 customers. During this period, the firm developed 140 distinct marketing campaigns, each defined by its objective, creative design, and eligibility rules (e.g., some targeted dormant users, whereas others focused on recent adopters of a specific tool or service). Each campaign could be deployed multiple times as weekly interventions, implemented in the form of randomized A/B tests. At the beginning of each week, the marketing department selected which campaigns from the portfolio of 140 would be activated. For each selected campaign, eligible customers were identified according to the campaign-specific qualification rule. Subsequently, each eligible customer was randomly assigned to either treatment or control.[1] Thus, control customers are those who qualified for a campaign in a given week but did not receive the intervention.

Throughout the paper, we distinguish between *campaigns* and *interventions*. A *campaign* refers to a unique marketing design defined by its objective, action, and reward. An *intervention* refers to the weekly execution of a campaign, during which eligible customers are randomly assigned to treatment or control. Thus, a single campaign can generate multiple interventions over time, depending on how frequently it is deployed. This distinction is important because it enables us to separate variation due to campaign design from variation driven by timing or customer composition, which is central to our decomposition of marketing effectiveness.

Similar to other firms in this sector, the focal company has a large customer base and engages in continuous promotional activity. On average, we observe nearly 70 *inter-*

[1] We corroborate that randomization at the intervention level was successful.

ventions per week, involving approximately 88,700 customers (treatment or control) each week. As described, each unique *campaign* design is implemented multiple times over the observation window. Figure 1 summarizes the size and treatment propensity of each campaign, aggregated across all weeks. The number of customers participating in a given campaign varies substantially, ranging from about 13,000 to over 263,000 (Figure 1a). On average, the allocation of eligible customers to the treatment condition is roughly 79% (Figure 1b).



(a) Number of customers per campaign.

(b) Proportion of customers in the treatment condition.

Figure 1: Campaign activity

The figure on the left shows the distribution of number of customers per campaign, including both treatment and control customers, across all campaigns in the data. The figure on the right shows the distribution of the proportion of customers in the treatment (versus control) condition.

Campaign Characteristics. Each campaign is defined by a set of features reflecting both managerial intent and executional design. Specifically, we classify campaigns along three main dimensions: `goal`, `action`, and `reward`.

The `goal` refers to the strategic objective of the campaign, such as reactivation (e.g., "wake-up"), cross-selling, or retention. The `action` dimension captures the behavior the customer must undertake to receive the benefit, typically involving account recharge or the purchase of a data pack. The `reward` specifies the type of incentive offered, most

commonly in the form of a credit balance (usable across services) or a promotional balance (restricted in usage). Figure 2 summarizes the campaign characteristics in our data.

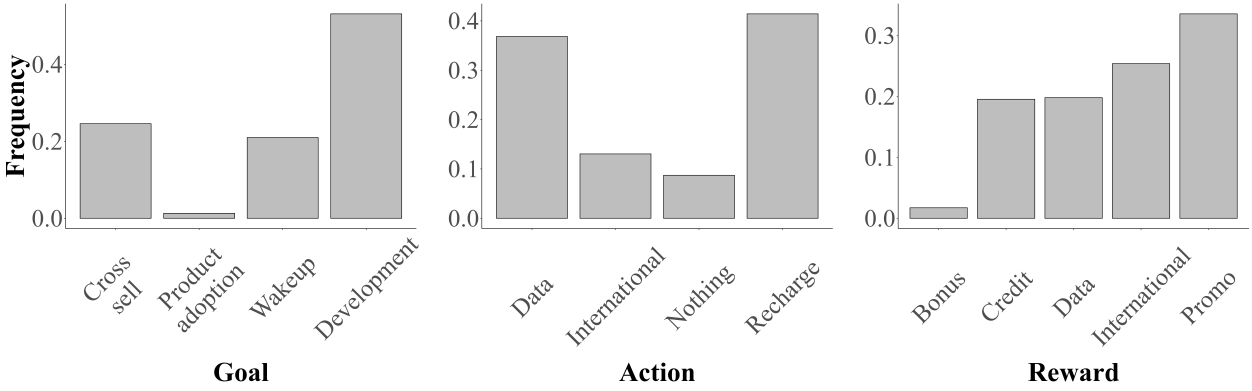


Figure 2: Types of campaigns
Proportion of (unique) campaigns with a particular Goal, Action, and Reward, respectively.

Because interventions are repeatedly drawn from a pre-defined portfolio (decided by the marketing team at the focal organization) and implemented across different points in time, the same campaign design appears multiple times in the dataset. This recurring structure allows us to observe treatment effect variation across campaign designs, time (capturing contextual factors), and customer segments. Table 1 provides the summary statistics across campaign designs.

Table 1: Campaign summary statistics.
Summary statistics for the campaign data. Campaign features are first aggregated at the campaign level, then summarized across campaigns.

	Avg.	Std.Dev.	Min.	Max.	N
# qualified users	88,707	58,255	13,016	263,262	140
% targeted	0.79	0.04	0.58	0.86	140
# weeks implemented	69.6	22.6	17.0	96.0	140

Campaigns also include a qualification rule that determines which customers (among the entire customer base) are eligible to be considered for treatment in a given week. These rules are defined by the firm and are based on observable behavior (e.g., no recharge activity in the past seven days). Importantly, once a customer qualifies for a campaign,

assignment to treatment or control is fully randomized (see more details in Section 4.2), enabling us to consistently estimate treatment effects.

Repeated Observations per Customer. A distinctive feature of this empirical setting is the repeated exposure of customers to randomized marketing interventions. Over the approximately 2-year observation window, each customer appears in multiple interventions, either in the treatment or control group.[2] Note that customers could not be assigned to more than one campaign in the same week. On average, a customer is associated with 28 such interventions, though this number varies substantially across individuals (Figure 3).

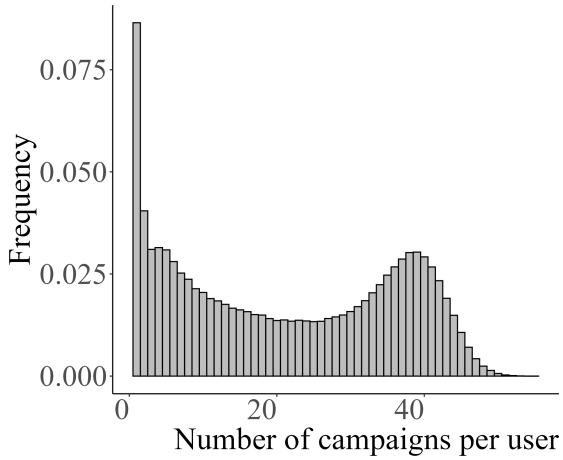


Figure 3: Number of interventions per user
Histogram of the number of interventions a customer participated in (either as treatment or control).

This repeated-measures structure is not uncommon in digital companies that frequently run A/B tests on their customers, and is central to our analysis. It enables us to track the same individual across different campaign designs, contextual conditions, and varying temporal distances from prior interventions. As a result, we observe substantial within-customer variation that can be exploited to estimate individual responsiveness

[2]Over time, customers can be exposed to the same or different campaign designs, depending on their qualification status.

and to evaluate dynamic patterns in treatment effects, including potential intervention saturation.

Because eligibility is determined week by week through predefined deterministic behavioral rules, and treatment is randomly assigned conditional on eligibility, we can isolate causal effects within each intervention while aggregating evidence across campaigns to assess the consistency of individual responses. This structure allows us to move beyond campaign-level analysis and quantify heterogeneity in responsiveness at the customer level, both within and across intervention types.

Customer Characteristics and Outcomes of Interest. Our primary behavioral outcome is the total monetary value recharged by a customer in the 30 days following an intervention, which reflects the focal firm's key performance indicator. To account for pre-existing differences in purchasing behavior, we also observe recharge value in the 30 days preceding the intervention and define our main dependent variable as the change in recharge value relative to this pre-intervention baseline.

For each customer–campaign instance, we observe a set of pre-treatment behaviors. Prior to each intervention, we capture customers' usage (in minutes) over multiple look-back windows (7, 14, and 30 days), which serve as proxies for recent engagement and potential eligibility. We also observe several features that characterize customers' promotional history. For each campaign, we know whether the customer was treated or held out, the timing of that exposure, and the cumulative history of campaign participation. These features (e.g., time since last campaign, cumulative number of prior campaigns, and recent usage) capture key dimensions of prior exposure and its relationship to subsequent responsiveness. We do not observe customers' demographic or socioeconomic information. Table 2 summarizes the distribution of these characteristics across customers.

Table 2: Customer summary statistics.
Summary statistics across customers. Variables are first aggregated at the customer level, then summarized across customers.

Variable	Mean	SD	Min	Max	N
Recharge next 30 days					
Average	87.56	127.71	0	3334	584,141
Coefficient of variation	1.05	0.79	0	8	
Usage past 7 days					
Average	23.72	32.78	0	559	584,141
Coefficient of variation	1.48	0.93	0	8	
Usage past 8–14 days					
Average	23.02	31.94	0	550	584,141
Coefficient of variation	1.52	0.91	0	8	
Usage past 15–30 days					
Average	50.49	68.18	0	1115	584,141
Coefficient of variation	1.15	0.73	0	8	
Tenure – # weeks since first intervention					
Average	28.18	15.94	1	84	584,141
Coefficient of variation	0.33	0.18	0	2	
Recency – # weeks since last intervention					
Average	3.41	8.13	1	119	570,731
Coefficient of variation	0.72	0.51	0	4	
Frequency – # interventions in the last 4 weeks					
Average	1.95	0.64	0	4	584,141
Coefficient of variation	0.72	0.25	0	3	

4. Modeling approach

Our objective is to leverage the structure of repeated randomized experiments to decompose treatment effect variation across three distinct dimensions: the design of the intervention, the timing of its delivery, and the responsiveness of individual customers. The central modeling challenge is to appropriately pool evidence across these multiple interventions while capturing observed and unobserved heterogeneity, systematic effects of campaign design, and contextual fluctuations over time.

4.1. Model Specification

To quantify the drivers of heterogeneity in marketing effectiveness, we specify a hierarchical Bayesian model (Rossi, McCulloch, and Allenby 1996) that decomposes the response to interventions into interpretable components at the customer, campaign, and temporal levels. The model captures the impact of marketing interventions (A/B tests) on customer demand (recharge amount) by decomposing marketing effectiveness into three sources of variation:

- **Customer-level effects**: These capture both observed and unobserved heterogeneity in individual responsiveness to marketing interventions. The observed component includes behavioral features such as recent usage patterns and prior exposure to campaigns (i.e., recency, frequency, and monetary value metrics). The unobserved component is modeled through random effects, allowing for persistent, time-invariant individual differences that are not captured by observable features. While analyzing campaigns one at a time allows heterogeneity to be captured based on observed customer characteristics, a unique aspect of our work is the ability to recover unobserved heterogeneity by jointly analyzing multiple experiments within a Bayesian framework.

- **Campaign-level effects**: These capture systematic variation across interventions. Each campaign is assigned a fixed effect that absorbs both its intrinsic appeal and the selection mechanics induced by qualification rules, which are themselves a function of prior behavior (e.g., time since last recharge).

- **Time-level effects**: These capture contextual variation common to all users and interventions, including temporal shocks (e.g., macroeconomic events, internal scheduling) and seasonality. Week-level fixed effects are included to control for such systematic temporal influences.

Formally, let y_{ij} denote the outcome for customer i in intervention j, and W_{ij} indicate treatment assignment. The model is specified as:

$$y_{ij} = \tau_{ij} W_{ij} + \boldsymbol{\beta}_i^\top \mathbf{X}_{ij} + \boldsymbol{\gamma}^\top \mathbf{Z}_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $\mathbf{X}_{ij}$ contains pre-treatment behavioral covariates (capturing baseline usage levels and trends in past consumption), and $\mathbf{Z}_{ij}$ includes proxies for past exposure to promotional activity as well as campaign and week fixed effects. Recall that interventions are weekly executions of campaign designs; hence, the fixed effects decompose variation across campaigns (design heterogeneity) and across interventions (temporal realizations of those designs). Appendix A presents a list of all variables entering the model specification.

The key quantity of interest is the treatment effect τ_{ij}, which varies across observations and is decomposed as:

$$\tau_{ij} = \bar{\tau} + \tau_{t(j)} + \tau_{c(j)} + \tau_{ij}^I, \tag{2}$$

where $\bar{\tau}$ is the population mean effect, $\tau_{t(j)}$ and $\tau_{c(j)}$ are week- and campaign-specific deviations, respectively, and τ_{ij}^I denotes the individual-level deviation, capturing systematic variation in treatment effects attributed to individual characteristics, both observed and unobserved. We further write

$$\tau_{ij}^I = (\boldsymbol{\gamma}^I)^\top \mathbf{V}_{ij} + u_i^I, \quad u_i^I \sim \mathcal{N}(0, \sigma_I^2), \tag{3}$$

where $\mathbf{V}_{ij}$ captures observed heterogeneity as well as exposure-history–dependent responsiveness[3], while u_i^I captures unobserved individual differences.[4]

[3] We also include the proportion of customers who where allocated to the treatment condition in that particular j.

[4] In this paper we use the term *dynamic* to refer to responsiveness that varies as a function of customers' prior exposure to marketing interventions (e.g., recency and cumulative exposure), rather than to forward-looking behavior or state-space dynamics.

The vector $\mathbf{V}_{ij}$ serves two related purposes. First, it captures heterogeneity associated with *observed* consumption histories (e.g., differences in baseline activity levels). Second, it allows treatment responsiveness to vary with prior exposure. To this end, $\mathbf{V}_{ij}$ includes: (i) discrete activity categories based on past 30-day usage quantiles, allowing for non-linear effects of prior engagement; (ii) *recency of prior treatment*, defined as the number of weeks since a customer last received any intervention, with a separate category for customers who have never been treated; and (iii) *treatment stock*, defined as a geometrically decayed sum of past treatments to capture cumulative exposure, with a weekly decay factor of 0.75.[5]

To maintain flexibility while keeping the model tractable, we discretize both recency and treatment stock. Recency is grouped into a small number of weekly bins, and treatment stock is discretized into quintiles. We further include interaction indicators between recency and stock, yielding a set of mutually exclusive exposure-history categories, along with a separate category for never-treated customers. Full details on bin definitions and construction are provided in Appendix A.

In sum, the model can be viewed as decomposing outcomes into three additive layers. First, baseline outcomes are explained by observed pre-treatment behavior and by campaign- and week-level fixed effects, which absorb systematic differences in demand across customer segments, campaign designs, and time periods. Second, the treatment effects τ_{ij} are allowed to vary flexibly across customers, campaigns, and weeks, with an overall average effect $\bar{\tau}$ augmented by campaign-specific and week-specific deviations, and an individual-level adjustment. Third, individual responsiveness itself is decomposed into an observed component, which depends on customers' prior usage and exposure history, and an unobserved component capturing persistent differences in sensitivity to marketing. Importantly, campaign and week effects enter the model twice. They first control for baseline outcome differences, and then allow treatment effects to vary by

[5]Note that to construct (ii) and (iii) we only used interventions in which the customer was in the treatment (as opposed to control) condition.

campaign and timing, ensuring that heterogeneity in effectiveness is identified separately from differences in underlying baseline outcomes.

4.2. Pooling Strategy

Our data comprise hundreds of interventions that vary in sample size, treatment–control allocation, and qualification rules. These features require a careful pooling strategy when estimating a single model across campaigns. Specifically, our approach addresses three issues that arise in this setting: Simpson's paradox, heterogeneity in campaign qualification rules, and overlap across campaigns.

Simpson's paradox. Simpson's paradox may arise when treatment effects are aggregated across interventions that differ in the composition of the underlying user population. In experimental settings, this concern is particularly relevant when campaigns span multiple time periods and treatment and control groups are exposed to systematically different mixes of users or contexts (e.g., weekdays vs. weekends, heavy vs. light users). In such cases, the aggregate treatment effect may differ from, or even contradict, the within-group effects (Crook et al. 2009). For example, a treatment may outperform control within each period but appear less effective in pooled data if treated users are disproportionately exposed during lower-performing periods.

To mitigate this concern, the level of aggregation in our model coincides with the unit of randomization, namely the intervention–week. We further include week-level fixed effects to absorb common contextual shocks and control for the proportion of users assigned to treatment within each intervention. This specification ensures that treatment effect estimation is not confounded by shifts in user composition or temporal variation, allowing us to separate persistent heterogeneity in responsiveness from transitory contextual effects (Kohavi and Longbotham 2023).

Campaign qualification rules. A second consideration for pooling arises from qualification-based targeting. Because campaigns differ in the behavioral criteria used to define eligibility not every customer is eligible for every intervention. For example, a "wakeup" campaign may target inactive users, whereas a "cross-sell" campaign may focus on recent rechargers. Importantly, eligibility is determined exclusively by observable pre-treatment behavior. That is, qualification is deterministic and known, and randomization into treatment or control occurs only after the eligible sample has been defined.

Because treatment assignment is randomized within the eligible population, qualification does not induce selection bias in the estimation of treatment effects. Conditional on eligibility, treatment assignment is unconfounded, and standard randomization-based inference applies. This feature, which is common in marketing interventions, does not compromise internal validity when estimating causal effects within the qualified population. Although qualification rules may limit the scope of counterfactual policy evaluation, they do not affect identification of heterogeneous treatment effects relative to control (Ellickson, Kar, and Reeder III 2023).

Accordingly, variation in qualification rules across campaigns does not interfere with our decomposition of treatment effect heterogeneity. We capture variation in qualification rules by campaign fixed effects, which capture both design-specific features and systematic differences in the composition of eligible users.

Overlap. A final consideration in our pooling strategy is whether stacking observations across campaigns is meaningful. If campaigns were targeted to largely disjoint customer subgroups, pooling would primarily combine unrelated sources of information, making pooling unattractive. In contrast, overlap in user participation links campaigns through shared customers, allowing estimation to leverage information across campaigns with comparable assignment mechanisms and exposures.

We assess this condition empirically using a graph-theoretic approach to examine the campaign-user structure. Specifically, we construct a *bipartite network*, a two-mode graph in which one set of nodes corresponds to campaigns and the other to users. An edge connects a campaign to a user if that user was eligible for (and randomized into treatment or control within) that campaign. This network captures the structure of campaign exposure over time and serves as the foundation for our overlap analysis.

To evaluate overlap across campaigns, we project the bipartite network into a one-mode campaign network. In this projection, two campaigns are connected if they share at least one user, with edge weights equal to the number of shared users.[6] The resulting design network is an undirected, weighted, complete graph with 140 vertices and 9,730 edges, where edge weights reflect the extent of overlap in campaign audiences. We analyze this network using standard descriptive measures from network analysis (Wasserman and Faust 1994).

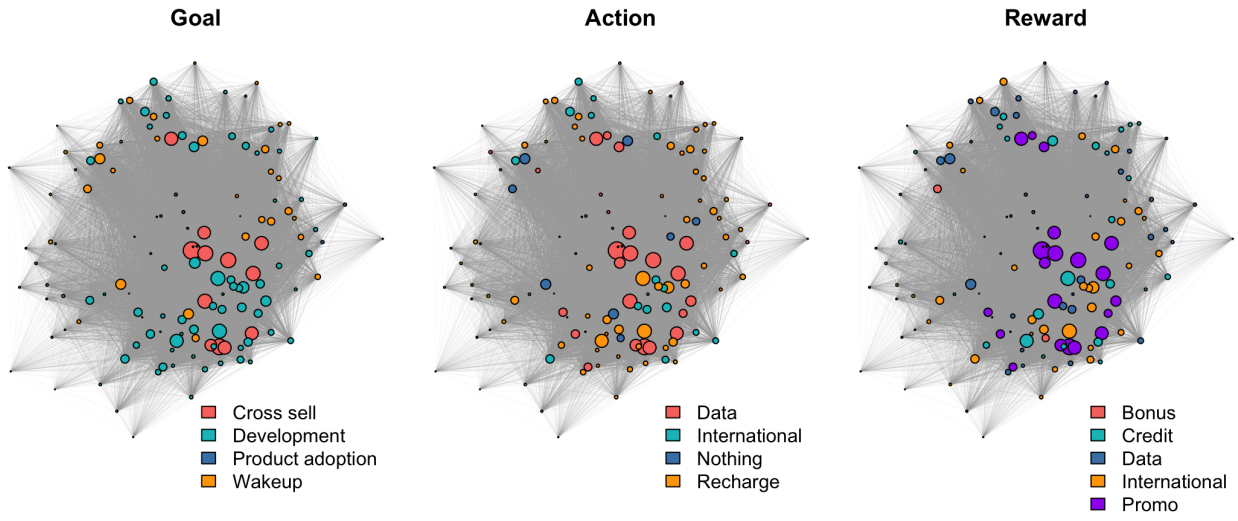


Figure 4: Campaign network visualization, colored by attribute
Campaign network visualization. Campaigns are connected to each other when they share the same customers, with edge weights capturing the amount of shared customers. Campaigns are colored by their Goal, Action, and Reward attributes, respectively.

[6]In principle, one could instead project the network onto users, but the resulting graph would be very large limiting its usefulness for the purpose of this campaing overlap analysis.

We begin by visualizing the design network, coloring campaigns by their `goal`, `reward`, and `action` attributes. As shown in Figure 4, nodes sharing the same attribute color do not cluster spatially. Were customers concentrated within subsets of similar campaigns, we would expect dense within-attribute connections and sparse cross-attribute links, yielding distinct color-based clusters in the network. We further quantify this lack of structure by examining assortativity and modularity with respect to campaign attributes. Partitions based on `goal`, `reward`, and `action` exhibit modularity scores close to zero (0.040, 0.000, and 0.010, respectively), indicating little alignment between network structure and campaign attributes. Together, these analyses indicate that customers do not cluster into isolated subsets associated with specific campaign types.

Finally, we apply multiple community detection algorithms, including Infomap (Rosvall and Bergstrom 2008), Label Propagation (Raghavan, Albert, and Kumara 2007), Louvain (Blondel et al. 2008), and Walktrap (Pons and Latapy 2005). Community detection tests whether the network exhibits any internally dense, externally sparse partition even in the absence of observable campaign attributes. Infomap and Label Propagation detect a single community, consistent with a densely connected network. Louvain and Walktrap identify three and two communities, respectively; however, these communities overlap substantially. Correlating community assignments with campaign attributes reveals no systematic alignment between detected communities and campaign attributes.

Overall, the campaign network exhibits extensive overlap across campaigns, with no evidence of segmentation by campaign attributes. This pattern indicates that campaigns are linked through shared customers rather than forming isolated groups, suggesting that stacking observations across the 140 campaigns is empirically meaningful. The proposed Bayesian model is designed to exploit this shared information.

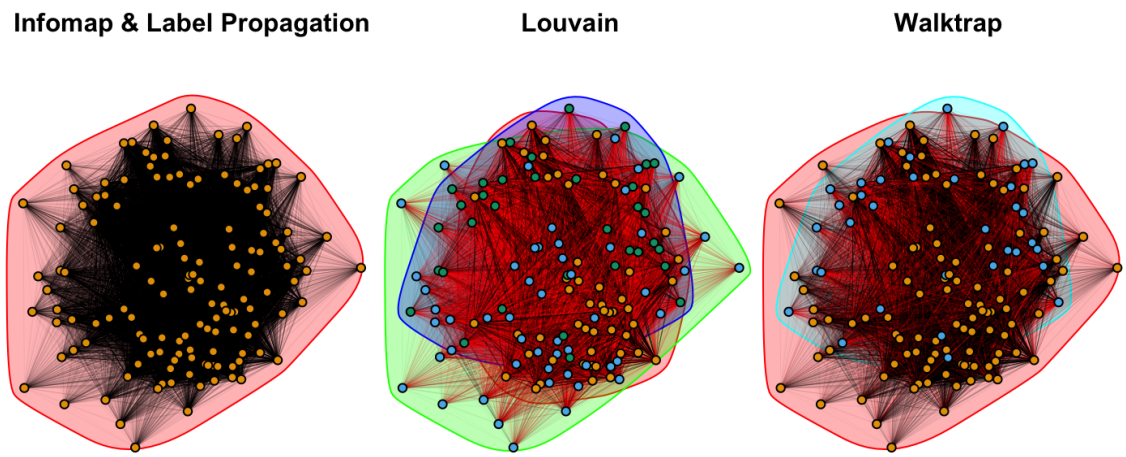


Figure 5: Community detection analysis, colored by community membership
Infomap and Label Propagation (left), Louvain (middle), and Walktrap (right). All algorithms find either one community or highly overlapping communities where campaign attributes are uniformly distributed across communities.

4.3. Estimation and Inference

We estimate the model using a fully Bayesian approach via Gibbs sampling. Each parameter of the model (see Equations 1 – 3) is assigned a standard conjugate prior, which enables efficient block-wise updates within each Markov Chain Monte Carlo (MCMC) sweep. After convergence of the MCMC chain, we sample the posterior distribution of the model parameters and base all subsequent analyses on these samples. Further implementation details are provided in Web Appendix A.2.1.

We randomly split the data (described in section 3) into a calibration and a holdout dataset. To create the holdout dataset, we randomly select one observation per user as holdout. We only consider users with at least 6 observations. Users with 5 observations or less will not appear in the holdout sample. The calibration sample has approximately 16 million rows and includes about 584K unique customers. The holdout dataset includes about 462K unique customers, and covers all 140 campaigns and all weeks that exist in the calibration data.

4.4. Examining the Predictive Ability of τ_{ij} Posteriors

Of central importance is the posterior distribution of τ_{ij}, the estimated treatment effect for individual i at occasion j. As specified in Equation 2, this quantity incorporates multiple sources of variation, including campaign-specific characteristics $c(j)$, broader contextual moderators such as week-level effects $t(j)$, and individual-level variation, both observed and unobserved. By jointly capturing these components, τ_{ij} summarizes treatment effect heterogeneity and forms the basis for our decomposition of variation across campaigns, weeks, and users.

Before decomposing this variation, we assess whether the inferred individual-level treatment effects (τ_{ij}) capture meaningful information about underlying heterogeneity in treatment response. To this end, we employ two complementary validation approaches. First, we compute the AUTOC (Area Under the Targeting Operating Characteristic; Yadlowsky et al. 2025) to evaluate whether ranking customers by predicted treatment effects improves treatment prioritization. Second, we implement the Best Linear Prediction (BLP) test, following Chernozhukov et al. (2025), to formally assess whether the model's predictions contain statistically detectable treatment-effect heterogeneity. Together, these tests evaluate both the statistical content of the estimated heterogeneity and its relevance for targeting decisions.

Area Under the Targeting Operating Characteristic. We begin by assessing whether the model's predicted treatment effects are informative for prioritization; that is, whether customers ranked as more responsive by the model indeed realize larger treatment effects. The AUTOC (Area Under the Targeting Operating Characteristic), introduced by Yadlowsky et al. (2025), provides a transparent metric for this purpose. AUTOC summarizes how much incremental treatment effect a firm would obtain by targeting customers in order of their *predicted* treatment effects $\hat{\tau}_{ij}$, relative to the average treatment effect (ATE) in a hold-out sample. If the model's ranking aligns with true treatment-effect het-

erogeneity, AUTOC is positive; if the predictions contain no informative signal, AUTOC is close to zero.

Formally, AUTOC is constructed from the Targeting Operating Characteristic (TOC), which evaluates, for every feasible targeting proportion (e.g., targeting the top 1%, 2%, . . .), the *realized* treatment effect among the highest-ranked customers. TOC values are expressed in the same units as the outcome (e.g., dollars of incremental revenue), facilitating direct economic interpretation. AUTOC aggregates these gains across all targeting proportions. Intuitively, TOC captures performance at each cutoff, while AUTOC summarizes overall targeting gains.

To account for estimation uncertainty, we compute AUTOC for each MCMC draw. For every posterior draw of $\hat{\tau}_{ij}$ in the hold-out sample, we construct the TOC curve by computing realized treatment effects for the top-ranked customers at each targeting threshold. We then compute the area under this curve and repeat the procedure across posterior draws and targeting thresholds, yielding a posterior distribution of AUTOC.

We find that the posterior distribution of AUTOC lies well above zero, indicating that ranking customers by predicted treatment effects leads to systematically higher realized gains in the hold-out sample. The posterior mean AUTOC is 1.92, with a 95% posterior interval of $[1.23, 2.48]$. In economic terms, this implies that, on average across targeting proportions, the treatment effect among targeted customers exceeds the population ATE by approximately \$2 per customer. Thus, the results show strong evidence that the model's predictions contain actionable signals.

Best Linear Prediction (BLP) Test. We complement the AUTOC analysis with a second validation exercise that directly tests whether predicted heterogeneity aligns with realized treatment effects. The Best Linear Prediction (BLP) test of Chernozhukov et al. (2025), adapted to randomized experiments by Athey, Keleher, and Spiess (2025), provides a parsimonious framework for this assessment. The intuition behind the BLP test

is straightforward. If the model's predicted effects $\hat{\tau}_{ij}$ contain genuine treatment-effect signal, then customers with higher predicted values should, on average, exhibit larger *realized* treatment effects in the hold-out data. The BLP regression formalizes this intuition by interacting each customer's treatment indicator with their predicted effect, and testing whether this interaction has a positive coefficient. A positive and significant interaction indicates that the model assigns higher scores to customers who benefit more from treatment.

We implement the BLP test in a fully Bayesian manner. For each MCMC draw, we estimate the BLP regression[7] on the held-out sample and record the posterior draw of the interaction coefficient. This yields a posterior distribution for the BLP slope, quantifying the strength and uncertainty of the relationship between predicted and realized treatment effects. The results confirm that the predictions encode meaningful causal signal. The 95% posterior interval for the BLP interaction coefficient is $[0.04, 0.21]$, with a posterior mean of 0.12, placing all posterior mass above zero. We therefore conclude that the model captures systematic variation in treatment effects across customers.

Together, the AUTOC and BLP tests provide complementary evidence that the model recovers economically meaningful and statistically detectable treatment-effect heterogeneity. This heterogeneity can be decomposed to address our central research questions regarding the relative roles of customer responsiveness, campaign design, and contextual timing in driving variation in treatment effects, and can also be leveraged to improve targeting performance, as we discuss next.

5. Model Insights

Having established that the estimated treatment effects τ_{ij} capture meaningful treatment-effect heterogeneity, we now examine how these effects vary systematically across cam-

[7] Equation (3.5) in Chernozhukov et al. (2025).

paigns, weeks, and individuals. Recall that τ_{ij} represents the estimated treatment effect for individual i at occasion j (Equation 2). This quantity incorporates multiple sources of variation, including campaign-specific characteristics associated with campaign $c(j)$; broader contextual factors such as events occurring in week $t(j)$; the customer's observed state V_{ij}, which summarizes prior consumption and recent marketing exposure; and individual-level unobserved heterogeneity u_i^I.

We organize the results by progressively examining these sources of variation. We begin with campaign-level heterogeneity, then consider week-level differences, and finally examine individual-level variation in treatment effects.

5.1. Campaign-Level Treatment Effect Variation

We first examine heterogeneity in average treatment effects across the 140 campaigns. Because each campaign is deployed repeatedly across multiple weeks, we compute a campaign-level average treatment effect by aggregating posterior draws of τ_{ij} across all calibration sample users and weeks associated with campaign c. This yields the posterior distribution of campaign-average effects, denoted $\bar{\tau}_{.c}$. Figure 6 presents the 95% posterior intervals for $\bar{\tau}_{.c}$, along with posterior means (dots), sorted from largest to smallest.

Posterior mean campaign effects range from approximately -2.8 to slightly above 3. The distribution of posterior means is skewed toward positive values, as reflected in the concentration of points to the right of zero. Accounting for posterior uncertainty, 10% of campaigns exhibit posterior intervals entirely above zero, whereas 7.1% lie entirely below zero. Negative effects are plausible in this setting, as some campaigns provide free minutes or credit, which can generate short-run cannibalization or post-promotional dips in spending.

Analyzing Campaigns in Isolation. One of the characteristics of our proposed approach of pulling information across campaign is that it should still preserve the origi-

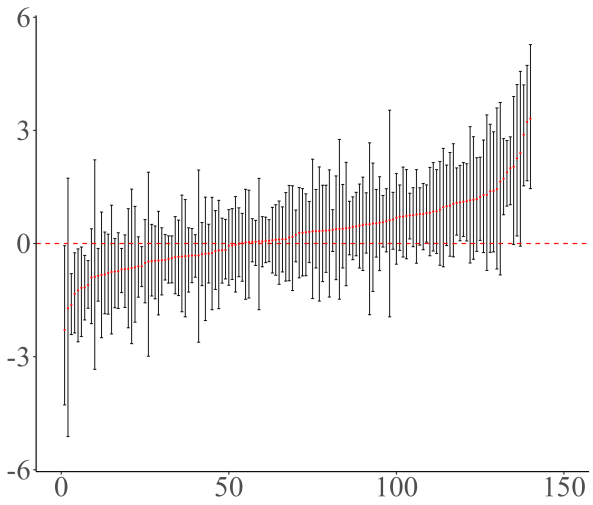
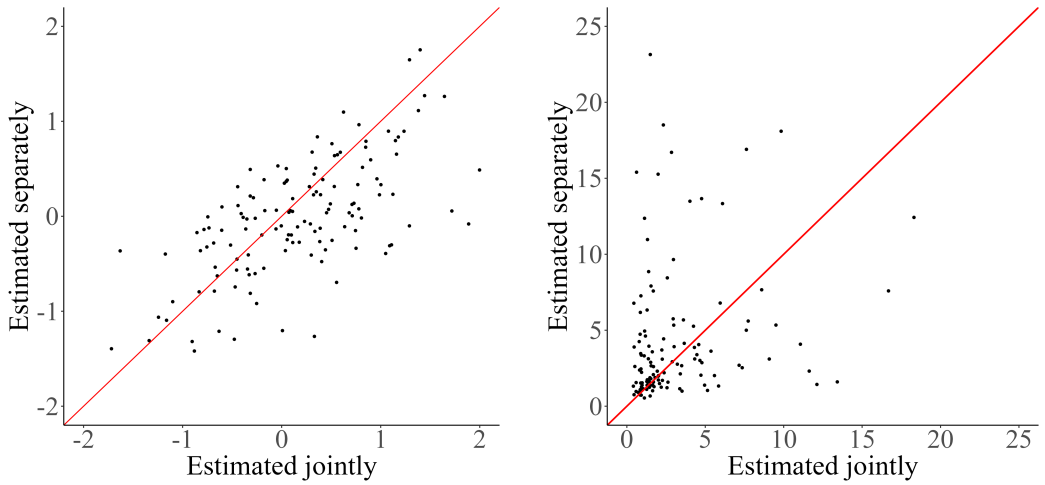


Figure 6: Posterior intervals for the 140 unique interventions
Posterior means (and 95% posterior intervals) for $\bar{\tau}_{.c}$, which corresponds to the average effect of each campaign c, integrated over the observed distribution of weeks and users.

nal within campaign average treatment effects (ATEs), since the within campaign effects should be valid independent of the pooling strategy. To test whether our pooling strategy preserves the original campaign ATEs we estimate a separate linear model for each campaign c, using the same set of control variables as in the joint model to ensure comparability.

Figure 7 contrasts ATEs obtained from the joint hierarchical Bayesian model with those obtained from campaign-by-campaign estimation. Panel (a) shows close alignment along the 45-degree line, indicating that both approaches recover similar point estimates for campaign-average effects. Panel (b), however, reveals systematic differences in precision. The coefficient of variation, measured as the ratio of the 95% posterior interval width to the posterior mean, is generally smaller under the joint model, as reflected by the concentration of points above the 45-degree line. This pattern indicates that pooling information across campaigns yields more precise estimates, with uncertainty that is smaller relative to effect magnitude.[8]

[8]This precision gain is also reflected in the share of campaigns with posterior mass away from zero. When estimated separately, approximately 2.5% of campaigns appear significantly negative and 2.5% significantly positive, as expected under sampling variability at the 2.5% and 97.5% tails. On the other hand,



(a) Average Treatment Effect. **(b)** Coefficient of Variation.

Figure 7: Comparison of ATEs estimated jointly vs. in isolation across campaigns
Scatter plots of posterior means and coefficient of variation of $\bar{\tau}_{.c}$ from the joint HB model (horizontal axis) and from campaign-by-campaign estimation (vertical axis). The 45-degree line is included for reference.

Taken together, these results show that our approach yields average treatment effects comparable to those obtained from campaign-specific analyses. However, analyzing A/B tests in isolation produces noisier campaign-level estimates. By jointly estimating all campaigns, the hierarchical model delivers meaningful gains in statistical efficiency, resulting in more precise estimates.

Effectiveness by Campaign Characteristics. Because we observe campaign attributes such as `goal`, `action`, and `reward`, we next examine whether these features are systematically associated with campaign effectiveness.

To motivate the analysis, it is helpful to clarify the interpretation of $\bar{\tau}_{.c}$. This parameter is obtained by aggregating the posterior draws of τ_{ij} (Equation 2) across all users and weeks for a given campaign c. Thus, $\bar{\tau}_{.c}$ reflects the *overall average effect* of campaign c on the outcome of interest. Importantly, its value is shaped not only by the campaign's intrinsic design but also by the weeks in which it was deployed and the users who partic-

under the joint model, 17.1% of campaigns exhibit posterior mass away from zero, reflecting increased power to detect meaningful effects.

ipated in each campaign-week. As a result, examining $\bar{\tau}_{\cdot c}$ in isolation would mix design effects with contextual and compositional factors.

Our model allows us to isolate these components. Specifically, we focus on the campaign-specific term $\tau_{c(j)}$ from Equation 2, which captures the intrinsic effectiveness of campaign design net of week- and user-level heterogeneity. We regress $\tau_{c(j)}$ on observed campaign characteristics.[9] This regression is estimated within each MCMC iteration, and we summarize posterior intervals for the coefficients (Web Appendix C.1, Table App-2). We find that "wake-up" interventions are significantly more effective relative to the average of other campaign goals: `Goal='WakeUp'` has a posterior mean of 0.22 with a 95% interval of [0.05, 0.39].[10] On the other hand, campaigns requiring international actions reduce spending: `Action='International'` has a posterior mean of -0.26 with interval $[-0.43, -0.10]$. Other campaign attributes do not exhibit systematic associations with effectiveness.

5.2. Week-Level Treatment Effect Variation

We next examine systematic variation in treatment effects across weeks. To this end, we analyze the posterior distribution of $\tau_{t(j)}$, corresponding to the week fixed effects in Equation 2. This component captures variation in treatment effectiveness that is uniquely attributable to week-to-week differences. As with $\tau_{c(j)}$, $\tau_{t(j)}$ reflects temporal variation after controlling for the specific campaigns deployed in each week and for the composition of exposed customers, as summarized by their observed states of past consumption and recent campaign activity.

[9] An alternative approach would be to include these characteristics directly as covariates in the structural model and treat residual campaign effects as random. We instead estimate campaign "fixed effects", which are precisely identified given repeated observations of each campaign, and relate observed attributes to effectiveness using posterior summaries.

[10] Consistent with this result, the data provider identified these campaigns, typically aimed at near-dormant users, as their most effective interventions.

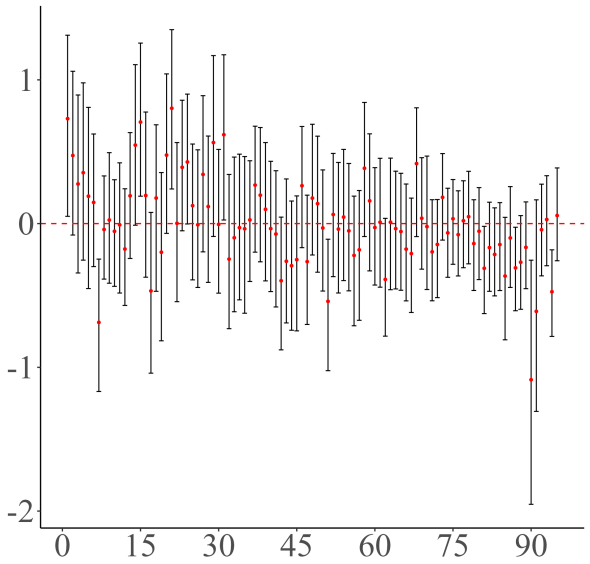


Figure 8: Posterior intervals across 95 weeks
Posterior means and 95% posterior intervals for $\tau_{t(j)}$, representing the variation in marketing effectiveness that is explained by the time component; that is, once the variation due to campaign design and customer heterogeneity has been accounted for.

Overall, we find limited evidence of substantial week-to-week variation in treatment effectiveness. Posterior means of $\tau_{t(j)}$ vary modestly, ranging from approximately -1 to 0.8, with most intervals overlapping zero (Figure 8). To test for a systematic trend, we regress $\tau_{t(j)}$ on time t within each MCMC draw. We find that campaign effectiveness has diminished slightly over time, although the effect is small in magnitude: the posterior mean of the slope for t is -0.005 (95% posterior interval $[-0.007, -0.003]$). Taken together, our findings suggest that aggregate market conditions and competitive dynamics were relatively stable over the observation period.

5.3. User-Level Treatment Effect Variation

A central objective of this paper is to characterize heterogeneity in individual responsiveness to marketing interventions. Because we observe users across multiple interventions over nearly two years, with repeated exposure to both treatment and control conditions, it is in principle possible to estimate an individual-level causal treatment effect. Natu-

rally, these estimates are subject to considerable uncertainty, given the relatively limited number of observations per user compared with other model components. Nevertheless, as we demonstrated in section 4.4, the individual-level effects may still reveal meaningful heterogeneity. We therefore examine user-level heterogeneity by decomposing it into observed heterogeneity (due to user characteristics), unobserved heterogeneity (systematic differences not captured by observables), and dynamics in responsiveness (possible saturation) arising from repeated exposure of marketing campaigns.

Moderating Role of Past Consumption (observed heterogeneity). We begin by investigating how users' prior consumption moderates their responsiveness to subsequent interventions. To avoid imposing a parametric relationship between past usage and treatment effects, we discretize past consumption and estimate the contribution of each usage level to individual responsiveness (captured in γ^I, Equation 3). This approach allows for flexible, potentially non-linear moderation by past consumption.

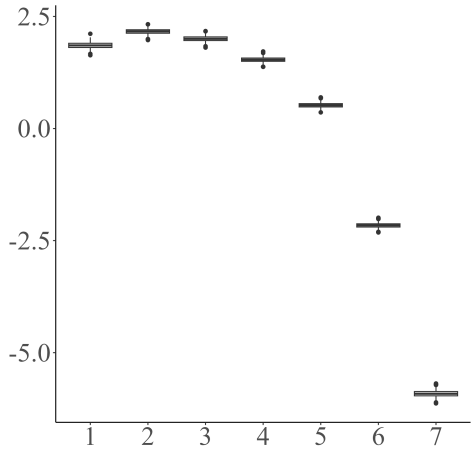


Figure 9: Posterior estimates of user-level responsiveness by levels of pre-treatment usage

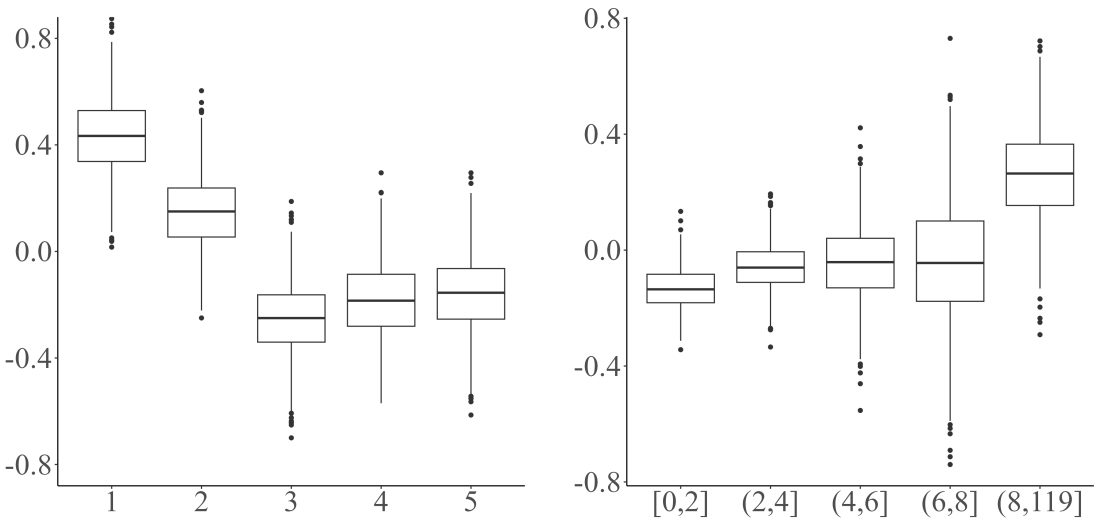
Posterior means and 95% intervals for coefficients γ^I across past-usage groups. Group 1 corresponds to the lowest-usage users (median $=$ 0 minutes in past 30 days), while group 7 corresponds to the highest-usage users (median $=$ 381 minutes in past 30 days).

The results in Figure 9 show non-linear observed heterogeneity, with treatment effects ranging from approximately -5 to 2.5 across usage groups. For users in the lowest usage

group (median 0 minutes in the prior 30 days), treatment effects are positive, whereas for users in the highest usage group (median $\approx$ 381 minutes), treatment effects are clearly negative. This pattern suggests that interventions stimulate consumption among low-engagement users but may cannibalize revenue among high-usage users, for whom promotions may substitute for spending that would have occurred in the absence of treatment. The non-linear relationship underscores the importance of flexible specifications for observed covariates.

Dynamics in Sensitivity to Interventions. Next, we examine how individual responsiveness varies with repeated exposure to marketing interventions and with the time elapsed since the most recent contact. Recall that we allow for non-linear exposure-history effects by discretizing both cumulative treatment exposure (treatment stock) and recency of past interventions, and estimating the contribution of each category to the individual treatment effect. Details on variable construction are provided in Web Appendix A.1.

Because responsiveness depends on interactions between recency and cumulative exposure, we report marginal effects that average predicted treatment effects over the empirical distribution of the interacting dimension. This approach isolates the partial relationship between responsiveness and each exposure-history component. Figure 10 presents posterior distributions of the marginal effects associated with treatment stock and recency, derived from the coefficients in γ^I. Two patterns emerge. First, responsiveness declines with cumulative exposure: users with higher treatment stock exhibit attenuated treatment effects (Figure 10a). This pattern is consistent with intervention fatigue, whereby repeated exposure reduces sensitivity to marketing contacts. Second, responsiveness increases with time since the last intervention (Figure 10b), although recovery is gradual and incomplete. In practice, more than eight weeks without exposure are required for responsiveness to approach its baseline level. Together, these results pro-



(a) Treatment stock **(b)** Weeks since last intervention

Figure 10: Posterior estimates of user-level responsiveness by recency and treatment stock

Posterior means and 95% credible intervals of treatment effects as a function of cumulative treatment stock (panel a) and recency (panel b). Effects are computed by averaging over the recency–stock interaction terms using the empirical distribution of the omitted dimension. For treatment stock (Figure a), group 1 corresponds to the users with the lowest treatment stock ($\approx$ 1.38 on average), while group 5 corresponds to the users with the highest treatment stock ($\approx$ 2.77 on average).

vide evidence of intervention fatigue with slow recovery, highlighting the importance of pacing interventions to sustain long-run effectiveness.

Unobserved Heterogeneity. Next, we examine the parameter u_i^I (Equation 3), which captures persistent, unobserved individual-level variation in responsiveness after accounting for campaign design, week effects, eligibility, and observed customer characteristics. In this sense, u_i^I represents intrinsic differences in responsiveness across customers. Figure 11 displays the posterior intervals of u_i^I for 250 randomly selected users.

The distribution of u_i^I reveals substantial unobserved heterogeneity in customer responsiveness. A non-trivial share of customers (13.56%) exhibit posterior intervals entirely above zero, indicating positive treatment effects even after conditioning on all other model components, including campaign design, temporal effects, and observed covariates. These users are systematically more responsive to interventions and generate in-

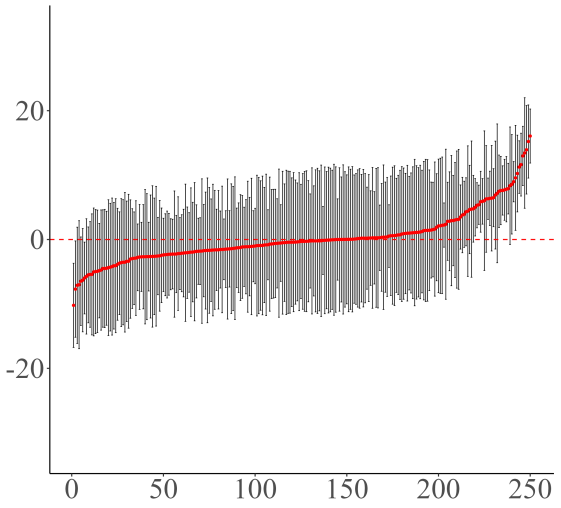


Figure 11: Posterior estimates of user-level responsiveness
Posterior means and 95% posterior intervals for u_i^I for 250 randomly selected users, representing unobserved heterogeneity in campaign responsiveness.

cremental revenue as a direct consequence of treatment. From a managerial perspective, this group represents a segment with persistently positive responsiveness that can be prioritized for promotional targeting. Conversely, the share of customers with posterior intervals entirely below zero is small (less than 1%), suggesting a limited subset of users for whom interventions are counterproductive. Such behavior may reflect strategic responses, such as delaying purchases in anticipation of promotions. At the same time, repeated targeting may itself induce strategic behavior, altering responsiveness over time. We examine these dynamic effects next.

6. Implications for targeting

Having established the presence of meaningful variation in treatment effects, we assess its managerial relevance by evaluating whether this heterogeneity can improve targeting decisions. We conduct a policy evaluation comparing a strategy that, for a given intervention, prioritizes customers based on posterior draws of τ_{ij} with a benchmark random allocation from the set of eligible customers. Our counterfactual targeting procedure pro-

ceeds as follows. All calculations are performed for each MCMC iteration using hold-out data only.

1. Calculate $\hat{\tau}_{ij}$ for each holdout observation.
2. Construct two policies:
   - **Model-based policy**: target users with a positive value of $\hat{\tau}_{ij}$.
   - **Random policy**: target the same number of users as under the model-based policy, but selected uniformly at random from the eligible set.
3. For each policy compute the expected recharge amount over the subsequent 30 days using realized outcomes and an inverse probability weighting (IPW) estimator.[11]

This design yields a sample of the posterior distribution of expected outcomes and policy gains. It evaluates whether the model's estimates of heterogeneous treatment effects translate into systematically better outcomes than random targeting. Because the evaluation is conducted on hold-out data, it isolates predictive signal from estimation noise. Importantly, the exercise does not alter the firm's qualification rules: all policies are evaluated within the originally eligible customer sets, mirroring the operational decision of whom to target conditional on eligibility.

Overall targeting value Table 3 reports posterior summary statistics of expected outcomes and the proportion of customers targeted under each policy, alongside the revenue observed under the firm's current policy (average recharge over the subsequent 30 days). The model-based policy yields higher expected revenue than both the firm's policy and the random benchmark, with posterior mass strictly above both alternatives. On average

[11] We note that our policy evaluation for both policies fully respects the firm's qualification rules: we only consider customers who satisfy the company-defined eligibility criteria when computing the IPW estimator for each campaign.

across all eligible observations, the model-based policy increases revenue by approximately 1.1% (including both targeted and non-targeted users) while reducing the share of customers targeted from about 80% to 50%.

Table 3: Policy evaluation results

Posterior means (and 95% posterior intervals) for the IPW estimator for revenue per customer and average proportion of customers targeted across policy approaches.

Policy	Revenue			Increase (%)	% Targeted
	Posterior Mean	(2.50% – 97.50%)		over Company Policy	average across interventions
Company	91.60	–	–	–	79.51%
Random	91.77	91.76	91.77	0.19%	49.49%
Proposed	92.61	92.24	92.97	1.10%	49.49%

From a managerial perspective, two implications follow. First, the model-based policy delivers a proportional revenue increase. Although a 1.1% average gain may appear modest in percentage terms, it corresponds to roughly $1 per customer per intervention and accumulates meaningfully given the firm's scale and repeated campaign activity. Second, the model-based policy targets substantially fewer customers than the firm's existing policy. This pattern suggests that the firm's baseline approach over-targets customers who are unlikely to benefit from intervention.

Importantly, the gains from the model-based policy arise not only from contacting fewer customers, but from identifying which customers to exclude from targeting. This distinction is underscored by the comparison with the random policy, which targets the same number of customers as the model-based policy but yields lower expected revenue. Moreover, reducing the share of customers targeted (from 80% to 50%) has additional benefits beyond short-run revenue. As reflected in our analysis of dynamic effects related to prior exposure (Section 5.3), targeting fewer customers mitigates intervention fatigue and, more broadly, reduces the risk of conditioning customers to expect frequent promotional outreach.

While the average gain from targeting is 1.1%, we find substantial differences in terms of the value of targeting across campaign types. To this end, we summarize policy per-

formance by campaign `Goal`, `Action`, `Reward`, and by the observed treatment propensity in the original randomized experiments. Full results are reported in Web Appendix C.2 (Table App-3).

Several patterns are noteworthy. First, gains are larger for behaviorally oriented campaigns. The model-based policy improves revenue by 2.5% for cross-sell/adoption campaigns and by 3.9% for wake-up campaigns, while it yields essentially no improvement for development campaigns, which are likely aimed at a longer time horizon. Second, gains are weaker for campaigns offering monetary credits or bonuses. Indeed, because we do not observe incentive amounts or associated costs, this information cannot be incorporated into the policy evaluation, limiting our ability to distinguish profitable from unprofitable targeting in these settings. Third, the observed treatment propensity in the original A/B test is a meaningful predictor of the gains achievable through targeting. A more balanced share of treatment and control is associated with higher value of targeting, reaching up to 10.8% improvement relative to the firm's policy. This pattern is intuitive: when users are treated most of the time, as the firm frequently did, limited exposure to control conditions constrains the model's ability to learn individual-level responsiveness to the campaign. While this mechanism is statistical in nature, it has important managerial implications. In the the explore-exploit experimental framework, firms that overly exploit (targeting most of their customers), may lose the ability to learn valuable treatment-effect heterogeneity for targeting. Instead, firms should preserve sufficient variation in treatment assignment, ensuring that both treated and control observations remain informative.

Discussion. It is important to clarify the scope of this "counterfactual" exercise. Because our evaluation respects the firm's original eligibility rules, we cannot assess how outcomes would change if eligibility criteria themselves were altered. Moreover, the policy evaluation is static and does not account for the dynamic effects of improved target-

ing on future consumption or promotion exposure, which we document in Section 5.3. Since the proposed policy targets substantially fewer customers each week, these estimates should therefore be interpreted as a conservative measure of the value of leveraging treatment-effect heterogeneity within the existing operational framework.

Taken together, our results indicate that the proposed targeting policy improves performance on average while delivering substantially larger gains in specific settings, particularly for non-monetary, behaviorally oriented campaigns and in experiments with sufficient randomization balance. In the next section, we quantify the contribution of customer, campaign, and contextual factors to improvements in targeting outcomes.

7. Decomposing Treatment Effect Heterogeneity of Marketing Interventions

Another central objective of this study is to disentangle the sources of heterogeneity in treatment effects. Having documented substantial variation, we now implement a decomposition that attributes this heterogeneity to customer, campaign-design, and contextual factors. We conduct two complementary analyses. First, we perform a variance decomposition to quantify the share of total variation in the estimated treatment effects τ_{ij} attributable to customer-, campaign-, and time-related components. Second, we assess managerial relevance by decomposing the overall targeting value of heterogeneity, as measured in the policy evaluation of Section 6, into the contributions of each component using Shapley values (Shapley 1953). Together, these analyses distinguish *where heterogeneity resides* from *where it contributes to actionable targeting improvements*.

7.1. Decomposing through Treatment Effect Variation in τ_{ij}

To quantify the relative importance of the sources of variation in Equation 2, we use the Bayesian hierarchical ANOVA framework of Gelman (2005) and Gelman and Hill (2006). For each component f (unobserved heterogeneity, observed heterogeneity, time, and campaign characteristics) we compute the finite-population standard deviation s_f of the corresponding term, where larger s_f indicates a larger contribution to dispersion in τ_{ij}. Intuitively, because the treatment effect τ_{ij} is an additive function of these components, a larger finite-population standard deviation s_f implies that component f generates greater dispersion in τ_{ij} across observations and therefore accounts for a larger share of treatment-effect heterogeneity. Following Gelman and Hill (2006) (pp. 460–462), we compute s_f for (1) unobserved customer heterogeneity u_i^I, (2) observed heterogeneity and dynamics $\gamma^I V_{ij}$, (3) time effects $\tau_{t(j)}$, and (4) campaign effects $\tau_{c(j)}$ in each MCMC draw by evaluating each component on the hold-out sample. To place components on a common scale, we then normalize contributions as

$$s_f^N = \frac{s_f}{\sum_f s_f}.$$

Technical details, including the treatment of random effects (u_i^I) versus fixed effects (all other components), are provided in Appendix B.1.

Table 4: Variance Decomposition.
Posterior means of estimated s_f's, and s_f^N and their 95% posterior intervals on percentage scale

Component	s_f	s_f^N	Posterior Interval (2.50% - 97.50%)	
Unobserved (u_i^I)	5.77	60.8%	60.0%	61.4%
Observed ($\gamma^I V_{ij}$)	2.84	29.9%	29.3%	30.3%
Time (week) ($\tau_{t(j)}$)	0.35	3.7%	3.2%	4.1%
Campaign ($\tau_{c(j)}$)	0.54	5.7%	5.1%	6.3%

Table 4 summarizes the posterior results. Customer-related components account for the vast majority of the variation in τ_{ij}: unobserved heterogeneity (u_i^I) accounts for 60.8%

and observed usage and dynamics ($\gamma^I V_{ij}$) for 29.9%. Campaign and week components account for comparatively little variation (5.7% and 3.7%, respectively). Overall, dispersion in treatment effects is primarily driven by differences across customers, consistent with prior evidence on the value of modeling individual-level heterogeneity (e.g., Rossi, McCulloch, and Allenby 1996).

Discussion. While informative, this variance decomposition has limitations. First, posterior uncertainty in individual-level effects (Figure 11) may mechanically increase the dispersion attributed to u_i^I. Second, the approach considers one term at-a-time while these terms may capture overlapping variation; for example, campaign effects may partly reflect eligibility rules that depend on past consumption. Third, the decomposition is statistical: it allocates dispersion across components but does not directly quantify contributions to outcome-relevant objectives such as incremental revenue or targeting performance.

7.2. Decomposing through Targeting Performance

To address the previous limitations, we conduct a complementary analysis using Shapley values. This approach allocates credit for model performance across potentially interacting sources of variation in outcome-relevant terms, thereby indicating not only where heterogeneity resides, but also where it is most actionable for targeting. Specifically, Shapley values address this attribution problem by defining each component's contribution as its *average marginal improvement* to a performance metric, taken over all possible sets in which components could be added to the model. Intuitively, the Shapley value asks: if we start from a baseline model and then sequentially add user-, campaign-, and week-level information in every possible combination, how much does each component increase targeting performance? This construction yields a fair allocation of the overall payoff across components while accounting for their joint presence in the model.

In our setting, we operationalize the payoff in two ways: (i) the expected policy gain from targeting users with $\hat{\tau}_{ij} > 0$ (IPW; Section 6) and (ii) the model's ranking performance (AUTOC; Section 4.4). We compute Shapley values for each payoff metric and attribute contributions to the user, campaign, and week components (details in Web Appendix B.2).

Because observed and unobserved heterogeneity are subcomponents of the user dimension, we report two complementary decompositions. First, we present the Shapley allocation across the three top-level components (user, campaign, and week). Second, we report a separate, within-user Shapley decomposition that allocates the user component between observed and unobserved heterogeneity. These within-user Shapley values are computed in a restricted game that conditions on the user dimension and therefore sum to 100%, not to the total-model Shapley shares.

Table 5: Shapley Values (as percentages) of Targeting Value metrics IPW and AUTOC.
Posterior means and 95% credible intervals of the Shapley value of each component.

	IPW		AUTOC	
Component	Shapley Value	95% interval	Shapley Value	95% interval
User (total)	88.8%	(78.3%, 99.1%)	88.8%	(78.3%, 94.7%)
Observed (conditional)	6.1%	(−14.9%, 19.2%)	12.6%	(2.4%, 21.4%)
Unobserved (conditional)	93.9%	(80.8%, 114.9%)	87.4%	(78.7%, 97.5%)
Campaign	9.0%	(2.3%, 16.1%)	9.1%	(4.4%, 18.2%)
Week	2.2%	(−4.2%, 8.9%)	2.0%	(−1.3%, 5.3%)

Two insights emerge. First, the Shapley allocations are similar across the two payoff metrics (IPW and AUTOC): in both cases, the user component accounts for the majority of targeting value, whereas campaign and week components contribute comparatively less. This indicates that customer-level differences are the primary driver of out-of-sample targeting gains. Second, within the user dimension, most of the targeting value is attributed to unobserved responsiveness captured by u_i^I. Combining the user share and the within-user unobserved share implies that approximately 83.4% ($= 0.888 \times 0.939$) of expected policy gains (IPW) are attributable to unobserved heterogeneity.

Overall, the Shapley decomposition is consistent with the variance decomposition in Table 4, while extending the analysis in two ways. First, by evaluating contributions in terms of targeting performance on hold-out data rather than posterior variance components, the Shapley decomposition directly considers economically relevant payoffs, with estimation noise entering only as attenuation. Second, the Shapley decomposition attributes value based on each component's marginal contribution when combined with the others, thereby capturing how campaign-, user- and week-level components operate jointly in generating targeting gains. The Shapley decomposition provides an outcome-focused summary of the sources of targeting performance improvements, highlighting the role of learning across campaigns in the presence of unobserved heterogeneity that cannot be recovered when campaigns are analyzed in isolation.

8. Conclusion and Discussion

Most experimental causal inference research to date has analyzed one experiment at a time. While this approach leads to valid incrementality estimates it does not leverage the insights and actions that can be gained from combining information across experiments. Specifically, this research leverages the information across experiments to investigate the sources of heterogeneity in the effectiveness of marketing interventions by developing and applying a hierarchical Bayesian framework that decomposes variation in treatment effects into three components: individual responsiveness, campaign design, and contextual timing. By isolating the contribution of each source, our approach moves beyond describing treatment effect heterogeneity to explaining its structure and linking that structure directly to targeting decisions and policy value.

We apply this framework to a large-scale empirical setting in the telecommunications industry, leveraging data from hundreds of randomized interventions delivered to over a half a million customers over a 2-year period. Our analysis reveals substantial individual-

level heterogeneity in responsiveness, much of which remains unobserved even after controlling for past consumption and prior exposure to marketing interventions. Importantly, we find it to be the biggest driver of variation in targeting effectiveness, that can only be inferred from analyzing repeated observations, which is an integral part of the Bayesian model we propose.

In addition, we find strong evidence of diminishing responsiveness: repeated exposure to marketing interventions reduces customer sensitivity, with only modest recovery over time. This pattern is consistent with intervention fatigue, whereby successive exposures yield diminishing returns. The evidence of diminishing responsiveness highlights a fundamental trade-off: while frequent interventions may yield short-term gains, they can also erode long-term engagement. Responsiveness should therefore be treated as a history-dependent quantity that varies with customers' prior exposure to marketing interventions, rather than as a fixed attribute. Marketers must accordingly pace and manage interventions deliberately.

Together, these findings highlight significant opportunities to improve marketing effectiveness by aligning targeting strategies with the evolving responsiveness of individual customers.

The broader implications of our findings are both operational and strategic. Operationally, the decomposition framework clarifies which customers and moments are most valuable for personalization. Strategically, it provides guidance on where firms should invest analytic and experimentation resources, i.e., customer, campaign design, or timing, based on their contribution to response heterogeneity.

Our results point to the potential of repeated experimentation, when carefully designed and combined across campaigns, as a substitute for more intrusive forms of customer data. While implementing our framework requires observing outcomes across multiple experiments, the role of this data is fundamentally different from that in traditional data-driven personalization approaches. Rather than continually accumulating

granular customer features or maintaining rich behavioral histories, repeated experimentation enables the inference of stable, low-dimensional latent traits that summarize unobserved responsiveness. Once these traits are estimated, firms can substantially reduce their reliance on persistent identifiers, raw behavioral logs, or external data sources, and in principle retain only aggregated or transformed representations. In privacy-constrained environments, structured experimentation therefore offers a pathway to personalization that emphasizes data minimization and transformation, preserving targeting power while limiting long-term dependence on detailed individual-level data.

Despite the model's flexibility and scalability, several limitations merit consideration. First, estimates of individual heterogeneity are subject to posterior dispersion, particularly for users with limited exposure histories. Future work may benefit from exploring structured priors or regularization methods (Bhadra et al. 2019) to stabilize these estimates, although such choices may introduce trade-offs with scalability. Second, we evaluate policies with inverse probability weighting (IPW). While unbiased, IPW can become highly variable when propensities are extreme (Robins, Rotnitzky, and Zhao 1994; Kang and Schafer 2007). In our application, this limitation is mitigated by complementing IPW with rank-weighted estimands (Yadlowsky et al. 2025) such as the Treatment-Optimality Curve (TOC) and its area (AUTOC), which evaluate the same policy performance more robustly and with lower variance. More generally, overlap/balancing and doubly robust estimators can stabilize policy-value estimation (Li, Morgan, and Zaslavsky 2018; Tan 2010; Bang and Robins 2005). Third, our current specification leverages behavioral data but does not incorporate richer customer-level information, such as attitudinal metrics, or other potentially relevant behaviors like interaction with customer service or the use of auxiliary services. Including such data, where available, may enhance the interpretability and accuracy of model predictions. Fourth, while our framework accounts for campaign-level heterogeneity using fixed effects, it does not explicitly model how specific campaign attributes, such as goal or reward type, contribute to differences in effective-

ness. Future research could extend the framework to include campaign-level covariates (similar to Huang, Ascarza, and Israeli 2024), thereby supporting inference and generalization across campaign designs. Finally, our analysis is limited to a single empirical setting in the telecommunications industry. While we expect that customer heterogeneity will also be a key driver of variation in other contexts, this remains an open question. Future research should investigate how these dynamics play out in alternative domains such as e-commerce, subscription services, or digital media.

In sum, our study highlights how repeated experimentation can be leveraged to better understand variation in marketing effectiveness. By showing not only where heterogeneity resides but also where it is actionable, we hope this work encourages firms to leverage repeated experimentation to uncover latent differences that improve targeting, and motivates future research that extends this framework across domains and decision settings.

References

Ascarza, Eva (2018), "Retention futility: Targeting High-risk Customers Might be Ineffective," *Journal of Marketing Research*, 55 (1), 80–98.

Ataman, M Berk, Harald J van Heerde, and Carl F Mela (2010), "The Long-term Effect of Marketing Strategy on Brand Sales," *Journal of Marketing Research*, 47 (5), 866–882.

Athey, Susan, Niall Keleher, and Jann Spiess (2025), "Machine Learning Who to Nudge: Causal vs Predictive Targeting in a Field Experiment on Student Financial Aid Renewal," *Journal of Econometrics*, 249, 105945.

Athey, Susan, Stefan Wager, and Vasilis Syrgkanis (2021), "Policy Learning with Observational Data," *Econometrica*, 89 (1), 133–161.

Bang, Heejung and James M Robins (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61 (4), 962–973.

Bell, David R, Jeongwen Chiang, and Venkata Padmanabhan (1999), "The Decomposition of Promotional Response: An Empirical Generalization," *Marketing Science*, 18 (4), 504–526.

Bhadra, Anindya, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard (2019), "Lasso Meets Horseshoe: A Survey," *Statistical Science*, 34 (3), pp. 405–427.

Bijmolt, Tammo HA, Harald J van Heerde, and Rik GM Pieters (2005), "New Empirical Generalizations on the Determinants of Price Elasticity," *Journal of Marketing Research*, 42 (2), 141–156.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008), "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10), P10008.

Chan, Tat, Chakravarthi Narasimhan, and Qin Zhang (2008), "Decomposing Promotional Effects with A Dynamic Dtructural Model of Flexible Consumption," *Journal of Marketing Research*, 45 (4), 487–498.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val (2025), "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India," *Econometrica*, 93 (4), 1121–1164.

Crook, Thomas, Brian Frasca, Ron Kohavi, and Roger Longbotham "Seven pitfalls to avoid when running controlled experiments on the web," "Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining," pages 1105–1114 (2009).

Dell'Acqua, Fabio, Maxime Cohen, and Yichuan Zhou (2021), "Personalization through Experimentation: Evidence from Targeted Promotions in a Two-Sided Market," *Management Science*.

Ellickson, Paul B, Wreetabrata Kar, and James C Reeder III (2023), "Estimating Marketing Component Effects: Double Machine Learning from Targeted Digital Promotions," *Marketing Science*, 42 (4), 704–728.

Ellickson, Paul B, Wreetabrata Kar, James C Reeder III, and Guang Zeng (2025), "Using Contextual Embeddings to Predict the Effectiveness of Novel Heterogeneous Treatments," *Available at SSRN 4845956*.

Gelman, Andrew (2005), "Analysis of Variance: Why It Is More Important than Ever," *The Annals of Statistics*, 33 (1), 1–31.

Gelman, Andrew and Jennifer Hill (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models* Analytical Methods for Social Research, Cambridge University Press.

Gordon, Brett R, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky (2019), "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," *Marketing Science*, 38 (2), 193–225.

Gupta, Sunil (1988), "Impact of Sales Promotions on When, What, and How Much to Buy," *Journal of Marketing Research*, 25 (4), 342–355.

Hitsch, Günter J, Sanjog Misra, and Walter W Zhang (2024), "Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation," *Quantitative Marketing and Economics*, 22 (2), 115–168.

Huang, Ta-Wei and Eva Ascarza (2024), "Doing More with Less: Overcoming Ineffective Long-term Targeting Using Short-term Signals," *Marketing Science*, 43 (4), 863–884.

Huang, Ta-Wei, Eva Ascarza, and Ayelet Israeli (2024), *Incrementality Representation Learning: Synergizing Past Experiments for Intervention Personalization* Harvard Business School.

Ibragimov, Marat, Artem Timoshenko, Duncan Simester, Jonathan A. Parker, and Schoar Antoinette (2025), "Improving Targeting Policies Using Transfer Learning," *Available at SSRN 5146292*.

Imai, Kosuke and Michael Lingzhi Li (2021), "Experimental Evaluation of Individualized Treatment Rules," *Journal of the American Statistical Association*, pages 1–15.

Kang, Joseph D. Y. and Joseph L. Schafer (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22 (4), 523–539.

Kohavi, Ron and Roger Longbotham "Online Controlled Experiments and A/B Tests," "Encyclopedia of Machine Learning and Data Science," pages 1–13, Springer (2023).

Kopalle, Praveen K, Carl F Mela, and Lawrence Marsh (1999), "The Dynamic Effect of Discounting on Sales: Empirical Analysis and Normative Pricing Implications," *Marketing Science*, 18 (3), 317–332.

Krefeld-Schwalb, Antonia, Eli Rosen Sugerman, and Eric J Johnson (2024), "Exposing Omitted Moderators: Explaining Why Effect Sizes Differ in the Social Sciences," *Proceedings of the National Academy of Sciences*, 121 (12), e2306281121.

Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky (2018), "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, 113 (521), 390–400.

Lodish, Leonard M, Magid Abraham, Stuart Kalmenson, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens (1995), "How TV Advertising Works: A Meta-analysis of 389 Real-world Split Cable TV Advertising Experiments," *Journal of Marketing Research*, 32 (2), 125–139.

Mela, Carl F, Sunil Gupta, and Donald R Lehmann (1997), "The Long-term Impact of Promotion and Advertising on Consumer Brand Choice," *Journal of Marketing research*, 34 (2), 248–261.

Mela, Carl F, Kamel Jedidi, and Douglas Bowman (1998), "The Long-term Impact of Promotions on Consumer Stockpiling Behavior," *Journal of Marketing Research*, 35 (2), 250–262.

Pons, Pascal and Matthieu Latapy "Computing Communities in Large Networks Using Random Walks," "Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20," pages 284–293, Springer (2005).

Raghavan, Usha Nandini, Réka Albert, and Soundar T Kumara (2007), "Near Linear Time Algorithm to Detect Community Structures in Large-scale Networks," *Physical Review E*, 76 (3), 036106.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89 (427), 846–866.

Rossi, Peter E, Robert E McCulloch, and Greg M Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15 (4), 321–340.

Rosvall, Martin and Carl T Bergstrom (2008), "Maps of Random Walks on Complex Networks Reveal Community Structure," *Proceedings of the National Academy of Sciences*, 105 (4), 1118–1123.

Sahni, Navdeep S (2015), "Effect of Temporal Spacing Between Advertising Exposures: Evidence from Online Field Experiments," *Quantitative Marketing and Economics*, 13 (3), 203–247.

Sahni, Navdeep S (2016), "Advertising Spillovers: Evidence from Online Field Experiments and Implications for Returns on Advertising," *Journal of Marketing Research*, 53 (4), 459–478.

Shapley, Lloyd S (1953), "Stochastic Games," *Proceedings of the National Academy of Sciences*, 39 (10), 1095–1100.

Shchetkina, Kseniya and Ron Berman (2024), "When is Heterogeneity Actionable for Personalization," *Available at SSRN 4561737* Available at SSRN: https://ssrn.com/abstract=4561737.

Tan, Zhiqiang (2010), "Bounded, Efficient and Doubly Robust Estimation with Inverse Probability Weights," *Biometrika*, 97 (3), 661–682.

van Heerde, Harald J, Peter SH Leeflang, and Dick R Wittink (2004), "Decomposing the Sales Promotion Bump with Store Data," *Marketing Science*, 23 (3), 317–334.

Wasserman, Stanley and Katherine Faust (1994), *Social Network Analysis: Methods and Applications* Structural Analysis in the Social Sciences, Cambridge University Press.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager (2025), "Evaluating Treatment Prioritization Rules via Rank-weighted Average Treatment Effects," *Journal of the American Statistical Association*, 120 (549), 38–51.

Yoganarasimhan, Hema, Ebrahim Barzegary, and Abhishek Pani (2023), "Design and Evaluation of Optimal Free Trials," *Management Science*, 69 (6), 3220–3240.

Zantedeschi, Daniel, Eleanor McDonnell Feit, and Eric T Bradlow (2017), "Measuring Multichannel Advertising Response," *Management Science*, 63 (8), 2706–2728.

Web Appendix

Learning from Many Experiments: A Hierarchical Bayesian Framework for Decomposing Marketing Treatment Heterogeneity

A. Model Specification and Estimation

This appendix provides the complete specification of the variables included in the hierarchical Bayesian model described in Section 4.1 as well as the details about the estimation procedure.

A.1. Full Set of Variables

Table App-1 describes the data used in the estimation, indicating the model component to which they belong (Section 4.1), along with their definition and operationalization.

A.2. Estimation implementation

This section provides additional details about the proposed hierarchical Bayesian model and outlines the Gibbs sampling procedure used to sample the joint posterior distribution of the parameters. Specifically, we expand upon the model formulation presented in the main text by describing the full set of closed-form conditional distributions for the Markov chain Monte Carlo (MCMC) algorithm. The Gibbs sampler leverages the conditional conjugacy of model components to facilitate efficient computation. Together, these procedures enable scalable inference across the large number of customers and repeated experiments in our data. Convergence diagnostics and robustness checks are reported in Web Appendix A.2.3.

A.2.1. Model details and Gibbs sampling

For ease of reference, we include the model equations from the main document (Section 4.1). *Notation.* To mirror the estimation implementation, we write inner products

Table App-1: Variables included in the model specification.

	Variable	Details
y_{ij}	Outcome	Recharge in next 30 days − recharge in prior 30 days
$\mathbf{X}_{ij}$	Usage past 7 days	Log transformation of usage (in minutes)
	Usage past 7–14 days	Log transformation of usage (in minutes)
	Usage past 14–30 days	Log transformation of usage (in minutes)
	Rolling past usage	Log transformation of the rolling average of minutes used in past 30 days
$\mathbf{Z}_{ij}$	A constant term	Column of ones
	Campaign fixed effects	Categorical indicators
	Week fixed effects	Categorical indicators
	All elements of $\mathbf{V}_{ij}$	
$\mathbf{V}_{ij}$	Recency × Treatment stock	Recency is discretized in six bins: $(0,2]$, $(2,4]$, $(4,6]$, $(6,8]$, $(8+)$, "never treated"
		Treatment stock is decayed at 0.75 per week; then discretized into quintiles
		Model includes 25 combinations and "never treated" as separate dummy variables
	Activity level	Seven categories based on past-30-day usage quantiles
	Proportion targeted	Share of customers targeted in campaign j

explicitly using transpose notation in this appendix; in the main manuscript we adopt a lighter notation for readability where appropriate. The model is specified as

$$y_{ij} = \tau_{ij} W_{ij} + \boldsymbol{\beta}_i^\top \mathbf{X}_{ij} + \boldsymbol{\gamma}^\top \mathbf{Z}_{ij} + \epsilon_{ij}, \tag{App-1}$$

$$\tau_{ij} = \bar{\tau} + \tau_{t(j)} + \tau_{c(j)} + \tau_{ij}^I, \tag{App-2}$$

$$\tau_{ij}^I = (\boldsymbol{\gamma}^I)^\top \mathbf{V}_{ij} + u_i^I, \tag{App-3}$$

where y_{ij} is the outcome for customer i in campaign j, W_{ij} indicates the treatment assignment, $\mathbf{X}_{ij}$ is a $k_1 \times 1$ vector containing pre-treatment behavioral covariates, $\mathbf{Z}_{ij}$ is a $k_2 \times 1$ vector containing proxies for past campaign activity together with campaign and week fixed effects, and $\mathbf{V}_{ij}$ is a $k_3 \times 1$ vector with proxies for past campaign activity. Furthermore, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $u_i^I \sim \mathcal{N}(0, \sigma_I^2)$, and $\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}} + \boldsymbol{u}_i^\beta$, with $\boldsymbol{u}_i^\beta \sim \mathcal{N}_{k_1}(0, \boldsymbol{\Sigma}_\beta)$. In terms of sample size, we have $i = 1, 2, ..., n$ customers. The number of campaigns for customer i is J_i. The total sample size is $N = \sum_i J_i$.

In our estimation, we allow for a non-zero covariance between the (upper-level) error terms u_i^I and $\boldsymbol{u}_i^\beta$. Furthermore, we include an overall intercept in the model (included in $\mathbf{Z}_{ij}$). In addition, following Dong and Wedel (2017), all intercepts belonging to a model component sum-to-zero for identification. For instance, the weekly treatment effects $\{\tau_t\}_{t=1}^T$ in (App-2) satisfy the constraint $\sum_{t=1}^T \tau_t = 0$. Similarly, the parameters in $\boldsymbol{\gamma}$ in (App-1) corresponding to the week fixed effects sum-to-zero, etc. Furthermore, we mean center the dependent variable and all quantitative covariates before estimation.

We keep the model tractable for large samples by specifying conjugate priors for the parameters, such that our model is fully Gibbs. Specifically, we take normal distributions for regression (intercept, slope) parameters and inverse-gamma/Wishart distributions for variance–covariance parameters. We assume no prior knowledge in setting the prior parameter values.

To develop the Gibbs sampler, we rewrite the model as an 'interaction' model. Specifically, we substitute equation (App-3) in (App-2), and the resulting equation with $\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}} + \boldsymbol{u}_i^\beta$ in (App-1), i.e.,

$$\begin{aligned} y_{ij} &= (\bar{\tau} + \tau_{t(j)} + \tau_{c(j)} + (\boldsymbol{\gamma}^I)^\top \mathbf{V}_{ij} + u_i^I) W_{ij} + (\bar{\boldsymbol{\beta}} + \boldsymbol{u}_i^\beta)^\top \mathbf{X}_{ij} + \boldsymbol{\gamma}^\top \mathbf{Z}_{ij} + \epsilon_{ij} \\ &= (u_i^I, (\boldsymbol{u}_i^\beta)^\top)(W_{ij}, \mathbf{X}_{ij}^\top)^\top \\ &\quad + (\boldsymbol{\gamma}^\top, (\bar{\boldsymbol{\beta}})^\top, \bar{\tau}, (\boldsymbol{\tau}^T)^\top, (\boldsymbol{\tau}^C)^\top, (\boldsymbol{\gamma}^I)^\top) \\ &\quad \times ((\mathbf{Z}_{ij})^\top, (\mathbf{X}_{ij})^\top, W_{ij}, (W_{ij}\mathbf{D}_{ij}^T)^\top, (W_{ij}\mathbf{D}_{ij}^C)^\top, (W_{ij}\mathbf{V}_{ij})^\top)^\top + \epsilon_{ij}, \end{aligned} \tag{App-4}$$

where $\mathbf{D}_{ij}^T$ is a design vector based on effect coding, picking the $t(j)$'th element from $\boldsymbol{\tau}^T$, with $\boldsymbol{\tau}^T$ denoting a $(T-1) \times 1$ vector of weekly treatment effect intercepts τ_t, with $\tau_T = -\sum \tau_t$. More specifically, for an observation (i, j) that falls in week $t(j) \in \{1, \ldots, T\}$, the effect-coded design vector $\mathbf{D}_{ij}^T \in \mathbb{R}^{T-1}$ is given by

$$\mathbf{D}_{ij}^T = \begin{cases} \mathbf{e}_{t(j)} & \text{if } t(j) \in \{1, \ldots, T-1\}, \\ -\boldsymbol{\iota}_{T-1} & \text{if } t(j) = T, \end{cases} \qquad \text{so that} \qquad (\boldsymbol{\tau}^T)^\top \mathbf{D}_{ij}^T = \tau_{t(j)}.$$

Here, $\mathbf{e}_s$ denotes the s-th standard basis vector in $\mathbb{R}^{T-1}$ and $\boldsymbol{\iota}_{T-1}$ is the $(T-1) \times 1$ vector of ones. $\mathbf{D}_{ij}^C$ and $\boldsymbol{\tau}^C$ are defined similarly.

To be consistent, W_{ij} is also effect-coded for estimation (i.e., $W_{ij} \in \{-1, 1\}$, where $W_{ij} = 1$ if observation (i, j) is in the treatment group and $W_{ij} = -1$ if it is in the control group, see also Dong and Wedel 2017).

We rewrite (App-4) in the following compact form:

$$y_{ij} = \underbrace{\boldsymbol{u}_i^\top \mathbf{R}_{ij}}_{\text{random (unit-specific) part}} + \underbrace{\boldsymbol{\theta}^\top \mathbf{F}_{ij}}_{\text{fixed part}} + \epsilon_{ij}, \tag{App-5}$$

where

$$\boldsymbol{u}_i = \begin{bmatrix} u_i^I \\ \boldsymbol{u}_i^\beta \end{bmatrix}, \qquad \mathbf{R}_{ij} = \begin{bmatrix} W_{ij} \\ \mathbf{X}_{ij} \end{bmatrix}, \tag{App-6}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \gamma \\ \bar{\boldsymbol{\beta}} \\ \bar{\tau} \\ \boldsymbol{\tau}^T \\ \boldsymbol{\tau}^C \\ \gamma^I \end{bmatrix}, \qquad \mathbf{F}_{ij} = \begin{bmatrix} \mathbf{Z}_{ij} \\ \mathbf{X}_{ij} \\ W_{ij} \\ W_{ij}\,\mathbf{D}_{ij}^T \\ W_{ij}\,\mathbf{D}_{ij}^C \\ W_{ij}\,\mathbf{V}_{ij} \end{bmatrix}. \tag{App-7}$$

We define the stacked dependent variable as

$$\boldsymbol{y} = (y_{11}, y_{12}, \ldots, y_{1J_1}, y_{21}, y_{22}, \ldots, y_{2J_2}, \ldots, y_{n1}, y_{n2}, \ldots, y_{nJ_n})^\top.$$

We stack the random-effects from (App-6) as

$$\boldsymbol{u} = (\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top, \ldots, \boldsymbol{u}_n^\top)^\top.$$

Furthermore, we define the following two sparse design matrices:

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{11}^\top \\ \mathbf{F}_{12}^\top \\ \vdots \\ \mathbf{F}_{1J_1}^\top \\ \mathbf{F}_{21}^\top \\ \mathbf{F}_{22}^\top \\ \vdots \\ \mathbf{F}_{2J_2}^\top \\ \vdots \\ \mathbf{F}_{n1}^\top \\ \mathbf{F}_{n2}^\top \\ \vdots \\ \mathbf{F}_{nJ_n}^\top \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{11}^\top & & & \\ \mathbf{R}_{12}^\top & & & \\ \vdots & & & \\ \mathbf{R}_{1J_1}^\top & & & \\ & \mathbf{R}_{21}^\top & & \\ & \mathbf{R}_{22}^\top & & \\ & \vdots & & \\ & \mathbf{R}_{2J_2}^\top & & \\ & & \ddots & \\ & & & \mathbf{R}_{n1}^\top \\ & & & \mathbf{R}_{n2}^\top \\ & & & \vdots \\ & & & \mathbf{R}_{nJ_n}^\top \end{bmatrix}, \tag{App-8}$$

where the block–diagonal matrix $\mathbf{R}$ stacks the transposed vectors $\mathbf{R}_{ij}$ in (App-6) and the matrix $\mathbf{F}$ stacks the transposed vectors $\mathbf{F}_{ij}$ in (App-7). Finally, let $\widetilde{\mathbf{y}}_{\text{fix}} = \mathbf{F}\,\boldsymbol{\theta}$ and $\widetilde{\mathbf{y}}_{\text{rand}} = \mathbf{R}\,\boldsymbol{u}$.

Full conditional distribution for $\boldsymbol{\theta}$. The full conditional distribution for $\boldsymbol{\theta}$ is a K_θ-variate normal distribution:

$$\boldsymbol{\theta} \,\big|\, \text{rest} \;\sim\; \mathcal{N}_{K_\theta}\big(\tilde{\boldsymbol{\mu}}_\theta,\; \tilde{\boldsymbol{\Sigma}}_\theta\big)\,,$$

with

$$\tilde{\boldsymbol{\Sigma}}_\theta = \Big(\mathbf{V}_{\theta 0}^{-1} + \sigma^{-2}\,\mathbf{F}^\top\mathbf{F}\Big)^{-1}, \qquad \tilde{\boldsymbol{\mu}}_\theta = \tilde{\boldsymbol{\Sigma}}_\theta\,\Big(\mathbf{V}_{\theta 0}^{-1}\boldsymbol{\mu}_{\theta 0} + \sigma^{-2}\,\mathbf{F}^\top\big(\mathbf{y} - \widetilde{\mathbf{y}}_{\text{rand}}\big)\Big)\,,$$

and $K_\theta = (k_1 + k_2 + k_3 + T + C)$. We set the prior mean $\boldsymbol{\mu}_{\theta 0}$ to zero and the prior variance as $\mathbf{V}_{\theta 0} = 1000 \times I_{K_\theta}$.

Full conditional distribution for $1/\sigma^2$. The full conditional distribution for the error precision σ^{-2} is a gamma distribution:

$$\sigma^{-2} \,\big|\, \text{rest} \;\sim\; \mathcal{G}\left(a_0 + \tfrac{N}{2},\;\; \Big[\,\tfrac{1}{b_0} + \tfrac{1}{2}\,(\mathbf{y} - \widetilde{\mathbf{y}}_{\text{rand}} - \widetilde{\mathbf{y}}_{\text{fix}})^\top(\mathbf{y} - \widetilde{\mathbf{y}}_{\text{rand}} - \widetilde{\mathbf{y}}_{\text{fix}})\,\Big]^{-1}\right),$$

where we set the prior shape a_0 equal to 0.1 and the prior scale b_0 equal to 10.

Full conditional distribution for $\boldsymbol{\Omega}_u^{-1}$. The random effects $\mathbf{u}_i$ defined in (App-6) follow a (k_1+1)-variate normal distribution centered at zero with variance-covariance matrix Ω_u. As mentioned earlier, we allow for a non-zero covariance between u_i^I and $\mathbf{u}_i^\beta$, so we have

$$\boldsymbol{\Omega}_u = \begin{bmatrix} \sigma_I^2 & \boldsymbol{\sigma}_{I\beta} \\ \boldsymbol{\sigma}_{\beta I} & \boldsymbol{\Sigma}_\beta \end{bmatrix},$$

where $\boldsymbol{\sigma}_{I\beta} = \boldsymbol{\sigma}_{\beta I}{}^\top$ are the covariances between u_i^I and $\mathbf{u}_i^\beta$. We take the Wishart prior $\boldsymbol{\Omega}_u^{-1} \sim \mathcal{W}(a_0,\ \mathbf{B}_0)$ with $a_0 = k_1 + 3$ and $\mathbf{B}_0 = I_{k_1+1}$. Let $S_u = \sum_{i=1}^n \boldsymbol{u}_i \boldsymbol{u}_i^\top$, then

$$\boldsymbol{\Omega}_u^{-1} \,\big|\, \text{rest} \ \sim\ \mathcal{W}(a_0 + n,\ \mathbf{B}_n)\,, \qquad \mathbf{B}_n = \left(\mathbf{B}_0^{-1} + S_u\right)^{-1}.$$

We use the scale-matrix parameterization of the Wishart distribution in the above notation.

Full conditional distribution for $\boldsymbol{u_i}$. Let $\mathbf{y}_i$ and $\widetilde{\mathbf{y}}_{\text{fix},i}$ be the J_i rows in $\mathbf{y}$ and $\widetilde{\mathbf{y}}_{\text{fix}}$ corresponding to customer $i = 1, 2, ..., n$. Furthermore, $\mathbf{R}_i$ stacks $\mathbf{R}_{ij}^\top$ over j for i. The full conditional distribution is

$$\boldsymbol{u}_i \,\big|\, \text{rest} \ \sim\ \mathcal{N}_{k_1+1}\left(\tilde{\boldsymbol{\mu}}_{u_i},\ \tilde{\boldsymbol{\Sigma}}_{u_i}\right),$$

with

$$\tilde{\boldsymbol{\Sigma}}_{u_i} = \left(\boldsymbol{\Omega}_u^{-1} + \sigma^{-2}\,\mathbf{R}_i^\top \mathbf{R}_i\right)^{-1}, \qquad \tilde{\boldsymbol{\mu}}_{u_i} = \tilde{\boldsymbol{\Sigma}}_{u_i}\left(\sigma^{-2}\,\mathbf{R}_i^\top\,(\mathbf{y}_i - \widetilde{\mathbf{y}}_{\text{fix},i})\right).$$

A.2.2. Implementation

The model is implemented in `R` (R Core Team 2025) using `c++` through the R package `Rcpp` (Eddelbuettel and François 2011), and leverages multithreading to parallelize updates of user-specific parameters. This parallelization is essential given the scale of the data.

More specifically, the Gibbs sampler for the results in the main document was run on an Amazon EC2 instance (`c7i.2xlarge`) with 8 vCPUs and 16GiB of memory (Amazon Web Services, Inc. n.d.). The total run time of our MCMC algorithm to sample the

posterior distribution of the model parameters in the main document lasts about 8 hours, which we find acceptable given the size of the data.

We update the $\boldsymbol{u}_i$'s in parallel as our n is large. We use four parallel workers, which balance efficiency and overhead. We follow Stubner (2024) to manage the random number generators across the parallel chains.

Furthermore, we implemented several pre-processing steps before running the Gibbs sampler. For instance, the matrix $\mathbf{F}$ is sparse (most elements are zero) which through the R package `RcppEigen` (Bates and Eddelbuettel 2013) can be stored efficiently in memory by using a specialized representation storing only the nonzero coefficients. In addition, we pre-computed matrices such as $\mathbf{F}^\top\mathbf{F}$ and $\mathbf{R}_i^\top\mathbf{R}_i$, and transposed large matrices $\mathbf{F}$ and $\mathbf{R}_i$ once before iterating through the Gibbs sampler.

We relied on the following core software and packages:

- `Ubuntu` 24.04.1
- `GNU C++` 17 (`g++` 13.2.0) and `OpenMP` 4.5
- `R` 4.4.2 (R Core Team 2024)
- `Rcpp` 1.0.13.1 (Eddelbuettel and François 2011)
- `RcppArmadillo` 14.2.0.1 (Eddelbuettel and Sanderson 2014)
- `RcppEigen` 0.3.4.0.2 (Bates and Eddelbuettel 2013)
- `dqrng` 0.4.1 (Stubner 2024)
- `BH` 1.84.0.0 (Eddelbuettel, Emerson, and Kane 2024)
- `data.table` 1.16.2 (Barrett et al. 2024)

A.2.3. MCMC convergence

We ran the sampler for 25,000 iterations, discarding the first 10,000 as burn-in. From the remaining iterations we retained 1,000 draws (every 15th) for posterior summaries. Figure App-1 shows the traceplot of the log-likelihood kernel of the main equation (Equa-

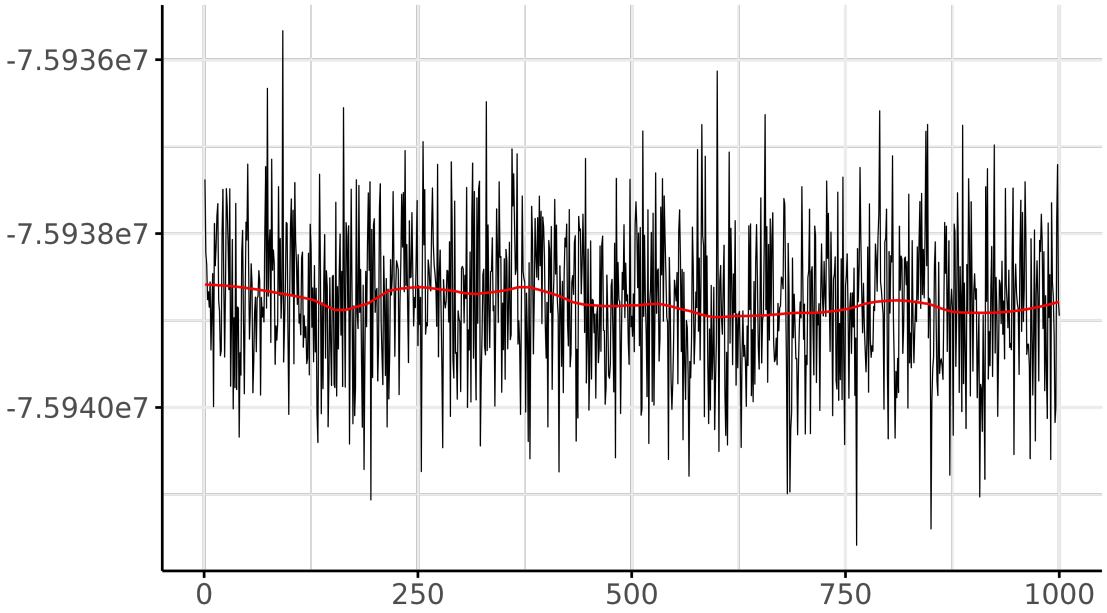


Figure App-1: MCMC traceplot log likelihood kernel
ESS=1000; AR(1)=0.02, AR(5)=0.04, AR(10)=-0.03.

tion 1). Autocorrelation among the retained draws is negligible (e.g., the lag-1 autocorrelation is 0.02), and the effective sample size is equal to the actual number of saved draws. Together, these diagnostics suggest good mixing and support the adequacy of the retained sample for posterior inference. Thinning was applied only to limit storage and post-processing overhead.

A.3. Exchangeability and overlap across interventions

Our hierarchical model pools information across interventions by assuming that, conditional on observed pre-treatment information, customers' remaining responsiveness parameters are drawn from a common distribution. It does not require that all customers be identical, but rather that, after conditioning on the information used to define eligibility and observed customer state, the u_i^I's are independent and identically distributed (Lindgren 1993; Gelman et al. 2000). This is a conditional exchangeability assumption.

Formally, the individual responsiveness component u_i^I enters the treatment-effect equation as

$$\tau_{ij} = \bar{\tau} + \tau_{t(j)} + \tau_{c(j)} + (\boldsymbol{\gamma}^I)^\top \mathbf{V}_{ij} + u_i^I, \qquad u_i^I \sim \mathcal{N}(0, \sigma_I^2).$$

Interpreted generatively, the prior $u_i^I \sim \mathcal{N}(0, \sigma_I^2)$ posits that the u_i^I are i.i.d. draws from a population distribution once we condition on the pre-treatment information that governs who is randomized within each intervention-week.[1] Importantly, these qualification rules enter the model in two places: in the outcome equation through pre-treatment covariates (capturing systematic behavioral differences across eligible groups) and in the treatment-effect equation through campaign and time components, thereby ensuring that pooling is conducted conditional on the segmentation structure that defines eligibility.

A practical concern is whether treatment-effect information learned in one set of campaigns is transferable to others. In the main text (Section 4.2), we address this empirically by documenting extensive overlap in customer participation across campaigns via the campaign–user design network (Figures 4 and 5). This overlap supports pooling under a common hierarchical prior by ensuring that campaigns are connected through shared customers rather than forming isolated subpopulations with no empirical bridge. When such overlap holds, posterior learning about the distribution of u_i^I and other components can propagate across interventions in a way that is grounded in observed cross-campaign participation.

[1] This conditioning information includes the firm's deterministic qualification rules (which are functions of pre-treatment behavior), as well as week effects and the observed customer state $\mathbf{V}_{ij}$.

B. Implementation Details for the Variance Decomposition

B.1. Gelman decomposition

We follow Gelman (2005) and Gelman and Hill (2006) to quantify the relative importance of different sources of variation in τ_{ij}. To quantify the sources of variation in ANOVA models, they estimate the finite-population standard deviations for batches of coefficients. How to estimate the standard deviations depends on whether the coefficients are 'modeled' (e.g., shrunk under an upper-level model) or '*un*modeled' (e.g., noninformative prior distributions). In order to quantify the relative importance of the sources of variation in τ_{ij}, i.e. campaign, week, observed and unobserved user-specific sources, we estimate the finite-population standard deviations for the corresponding terms in τ_{ij} following Gelman (2005) and Gelman and Hill (2006):

- for unobserved user-specific variation, we estimate the finite-sample standard deviation of u_i^I as $s_u = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(u_i^I - \bar{u}^I)^2}$, with $\bar{u}^I$ the sample mean of u_i^I, $i = 1, 2, ..., n$.

- for observed user-specific variation, we estimate the finite-sample standard deviation of past-usage (and similarly for recency and stock) as $s_p = \sqrt{\sum_{j=1}^{7}\lambda_j(\gamma^I_{p(j)} - \bar{\gamma}^I_p)^2}$, with λ_j the proportion of the sample in the j-th past-usage category (approximately $1/7$ by design), $p(j)$ the index corresponding to the j-the past-usage category coefficient in γ^I, and $\bar{\gamma}^I_p = \sum_{j=1}^{7}\lambda_j\gamma^I_{p(j)}$.

- for variation from campaign and week effects, we estimate the standard deviation in the same way as for observed user-specific variation.[2]

The estimates for the finite-population standard deviations s_f are computed on hold-out sample observations and in each sweep of the MCMC chain. The reported values in

[2]The finite-population standard deviation of the proportion of customers who received the intervention is estimated as the product of the standard deviation of the coin in the sample and the absolute value of the corresponding coefficient in $\boldsymbol{\theta}$ in Equation App-7.

Table 4 in the main document are posterior summaries. The relative importance, s_f^N, are computed by normalizing s_f.

B.2. Shapley decomposition

Our approach to decompose treatment effect heterogeneity is adapted from Hué et al. (2023). These authors introduce 'eXplainable PERformance (XPER)', a method to decompose model performance metrics into feature-level contributions. XPER is built on Shapley values (Shapley 1953). In game theory, Shapley values distribute payoffs among players. In XPER, performance metrics (e.g., Area Under the Curve (AUC), Mean-Squared Error (MSE), R-squared, etc.) are decomposed across model features. Each XPER value measures a feature's marginal contribution to performance by evaluating performance changes across different coalitions of features. XPER satisfies axioms associated with Shapley values (e.g., efficiency, symmetry, linearity, and null effect properties). This approach is both model-agnostic and metric-agnostic. Importantly, the authors propose an implementation approach that does not require re-estimating the models (with different sets of features).

In our application, each hierarchical dimension, customer, campaign, and time (week) heterogeneity, is treated as a "player" or "feature." We consider two payoffs that are relevant for policy evaluation: (a) the IPW of the policy $\tau_{ij} > 0$ (e.g., Section 6) and (b) the AUTOC based on τ_{ij} (e.g., section 4.4). For each feature f, the Shapley/XPER value φ_f measures the weighted average marginal contribution of feature f to the payoff measure (e.g., AUTOC) over all features coalitions. The marginal contribution of f is the difference between the expected value of the payoff (i) including f in and (ii) excluding f from the coalition. We compute the relative contribution of f by normalizing φ_f with respect to the difference between the payoff value and the payoff of a random targeting policy (φ_0). We compute these values in the holdout sample and in each iteration of the MCMC chain. Hence, the reported values in Table 5 in the main document summarize the posterior distributions of the Shapley/XPER values.

C. Further Results

C.1. Campaign Effectiveness by design characteristics

This appendix provides additional detail supporting the campaign-level results reported in the main text. In Section 5.1, we document substantial heterogeneity in average treatment effects across campaigns, summarized by the posterior distributions of campaign-level effects $\bar{\tau}_{\cdot c}$ (Figure 6). We further show that jointly estimating campaigns within the hierarchical Bayesian framework yields substantially more precise estimates than campaign-by-campaign analysis (Figure 7), and that only a limited subset of observable campaign characteristics is systematically associated with effectiveness.

Appendix C.1 extends these findings by reporting the full posterior regression results linking campaign-specific effects $\tau_{c(j)}$ to observable design characteristics, namely `Goal`, `Action`, and `Reward`. In addition, we provide complementary descriptive diagnostics that illustrate how campaign-level effects are distributed across design categories. Together, these results offer a more granular view of how campaign design features relate to intrinsic effectiveness, while reinforcing the conclusion from the main text that most cross-campaign variation remains unexplained by observable design attributes and instead reflects customer-level heterogeneity.

Figure App-2 presents posterior means and 95% credible intervals for campaign-level average treatment effects estimated *in isolation*. These estimates are shown for descriptive purposes only, to visualize the dispersion in effects across interventions without pooling information across campaigns. As discussed in Section 5.1, isolated estimation produces substantially noisier estimates relative to the joint hierarchical model, but the figure helps illustrate the wide range of effects observed across campaign designs.

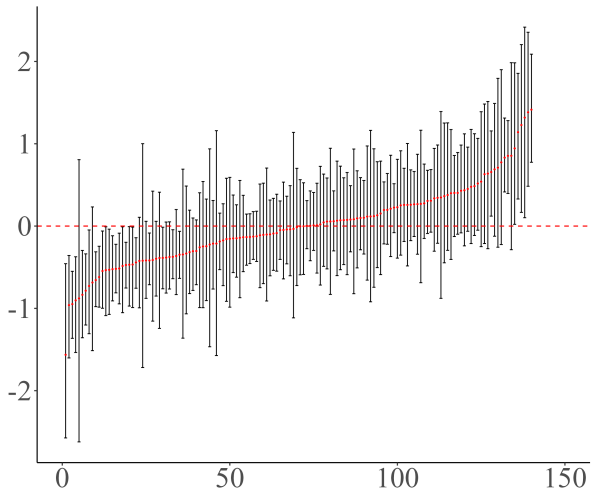


Figure App-2: Campaign-level ATEs with 95% intervals (estimated in isolation)
Posterior means and 95% credible intervals of campaign-level ATEs, estimated in isolation for descriptive comparison across interventions.

Table App-2 aggregates campaign-specific treatment effects by design characteristics, reporting posterior medians and 95% posterior intervals from the joint hierarchical Bayesian model. Consistent with the discussion in Section 5.1, campaigns with a `Wakeup` objective exhibit systematically higher effectiveness on average, whereas campaigns involving `International` actions or `Promo`-based rewards tend to perform worse. Importantly, however, the posterior intervals within each category remain wide, highlighting substantial residual heterogeneity across campaigns even after conditioning on design features.

Table App-2: Posterior intervals of campaign effectiveness by design factors

Posterior medians and 95% posterior intervals of treatment effects across 140 interventions, grouped by design factor.

Category	Factor	2.5%	Mean	97.5%
Overall		–0.22	–0.08	0.06
Goal	Cross sell	–0.39	–0.17	0.06
	Development	–0.21	–0.06	0.09
	Product adoption	–0.34	0.01	0.36
	Wakeup	0.05	0.22	0.39
Action	Data	-0.12	0.13	0.39
	International	–0.43	–0.26	–0.10
	Nothing	–0.16	0.02	0.19
	Recharge	–0.02	0.10	0.23
Reward	Bonus	–0.22	0.07	0.37
	Credit	–0.16	0.00	0.16
	Data	–0.08	0.06	0.19
	International	–0.03	0.11	0.24
	Promo	–0.55	–0.23	0.09

These results underscore that observable campaign characteristics explain only a portion of the variation in campaign effectiveness. While certain design choices are associated with higher or lower average effects, a large share of cross-campaign heterogeneity persists within design categories. This finding aligns with the main-text decomposition results, which show that customer-level heterogeneity is the dominant source of variation in treatment effects.

C.2. Targeting Performance by Campaign Characteristics

This appendix provides additional detail underlying the targeting results reported in the main text. In Section 6, we evaluate the overall performance of the proposed model-based targeting policy and show that, on average, it improves revenue while substantially reducing the share of customers targeted. We further document that these gains vary systematically across observable campaign characteristics. Appendix C.2 reports the full set of results supporting these patterns.

Specifically, we summarize targeting performance by campaign `Goal`, `Action`, `Reward`, and by the observed treatment propensity in the original randomized experiments. Table App-3 reports posterior means of expected revenue per customer and targeting intensity under the proposed model-based policy and the firm's existing policy. Results are aggregated by campaign characteristic using posterior draws from each MCMC iteration, thereby fully propagating uncertainty from the estimation of heterogeneous treatment effects to policy evaluation.

Several patterns emerge. For campaign `Goal`, the proposed policy yields improvements of 2.5% for cross-sell/adoption campaigns and 3.9% for wake-up campaigns, while delivering negligible gains for development campaigns. This pattern is consistent with the informational nature of development campaigns, whose primary objective is not to elicit an immediate behavioral response.

For `Action`, gains are concentrated in campaigns involving data usage and international actions, whereas recharge-only and "no action" campaigns exhibit little improvement. For `Reward`, the proposed policy underperforms the firm's policy for credit/bonus campaigns. Because monetary incentives may induce spending that would have occurred regardless, and because incentive amounts are unobserved, the model cannot condition targeting decisions on campaign cost-effectiveness in these settings.

Finally, treatment propensity plays a central role in determining targeting gains. When treatment assignment is highly unbalanced (average propensity above approximately 80%), the proposed policy performs worse than the firm's policy. In contrast, campaigns with more balanced assignment exhibit substantially larger gains. This pattern reflects an

Table App-3: Targeting performance by campaign characteristics

Proportional revenue gains from targeting based on model-predicted treatment effects relative to the company's existing policy. Results are reported by campaign characteristics, including Goal, Action, Reward, and treatment propensity (i.e., the proportion of customers assigned to treatment in the original A/B test). Values in the final column are shown in bold when the corresponding posterior interval excludes zero.

Campaign type	# Camp.	Revenue (Proposed)	Revenue (Company)	% Targeted (Proposed)	% Targeted (Company)	Increase (%)
Goal						
Cross-sell/Adoption	19	125.09	122.02	0.47	0.79	**2.51**
Development	86	115.42	115.37	0.49	0.80	0.04
Wake-up	35	22.78	21.92	0.53	0.80	**3.94**
Action						
Data	32	118.87	116.52	0.48	0.79	**2.02**
Recharge	84	84.17	84.37	0.50	0.80	−0.24
Nothing	8	21.71	21.74	0.52	0.81	−0.15
International	16	113.72	111.30	0.48	0.79	**2.17**
Reward						
Promo	20	126.46	125.77	0.47	0.80	0.55
Data	46	84.52	80.24	0.50	0.78	**5.34**
Credit/Bonus	35	63.29	64.33	0.51	0.82	**−1.62**
International	39	85.74	85.02	0.50	0.80	**0.84**
Propensity (% Treated)						
[0.575, 0.786)	47	91.50	82.56	0.51	0.75	**10.84**
[0.787, 0.805)	46	94.81	92.86	0.49	0.80	**2.10**
[0.805, 0.856]	47	91.56	93.98	0.50	0.82	**−2.58**

information constraint: when users are treated almost always, limited individual-level counterfactual variation restricts the model's ability to identify heterogeneous effects, even in large samples.

To assess whether treatment propensity is mechanically correlated with campaign characteristics, we conduct chi-square tests between propensity bins and campaign `Goal`, `Action`, and `Reward`. No significant correlation is found for `Goal` or `Action` (p-values $>$ 0.05). For `Reward`, however, the correlation is significant (p-value $<$ 0.01). Data-reward campaigns account for 57% of campaigns in the lowest-propensity category but only 12.8% in the highest-propensity category, while credit/bonus campaigns increase from 12.8% to 36% across these categories. This pattern indicates that monetary rewards are overrepresented among high-propensity interventions.

References

Amazon Web Services, Inc. "Amazon Elastic Compute Cloud (Amazon EC2)," https://aws.amazon.com/ec2/ (n.d.) Last accessed: 2025-09.

Barrett, Tyson, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, Toby Hocking, and Benjamin Schwendinger *data.table: Extension of 'data.frame'* (2024) https://r-datatable.com, r package version 1.16.2, https://Rdatatable.gitlab.io/data.table, https://github.com/Rdatatable/data.table.

Bates, Douglas and Dirk Eddelbuettel (2013), "Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package," *Journal of Statistical Software*, 52 (5), 1–24.

Dong, Chen and Michel Wedel (2017), "BANOVA: An R Package for Hierarchical Bayesian ANOVA," *Journal of Statistical Software*, 81 (9), 1–46.

Eddelbuettel, Dirk, John W. Emerson, and Michael J. Kane *BH: Boost C++ Header Files* (2024) https://github.com/eddelbuettel/bh, r package version 1.84.0-0, https://dirk.eddelbuettel.com/code/bh.html.

Eddelbuettel, Dirk and Romain François (2011), "Rcpp: Seamless R and C++ Integration," *Journal of Statistical Software*, 40 (8), 1–18.

Eddelbuettel, Dirk and Conrad Sanderson (2014), "RcppArmadillo: Accelerating R with high-performance C++ linear algebra," *Computational Statistics and Data Analysis*, 71, 1054–1063.

Gelman, Andrew (2005), "Analysis of Variance: Why It Is More Important than Ever," *The Annals of Statistics*, 33 (1), 1–31.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2000), *Bayesian Data Analysis* Boca Raton: Chapman & Hall/CRC.

Gelman, Andrew and Jennifer Hill (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models* Analytical Methods for Social Research, Cambridge University Press.

Hué, Sullivan, Christophe Hurlin, Christophe Pérignon, and Sébastien Saurin (2023), "Measuring the Driving Forces of Predictive Performance: Application to Credit Scoring," *HEC Paris Research Paper No. FIN-2022-1463*.

Lindgren, Bernard W. (1993), *Statistical Theory* Chapman Hall.

R Core Team *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria (2024) https://www.R-project.org/.

R Core Team *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria (2025) https://www.R-project.org/.

Shapley, Lloyd S (1953), "Stochastic Games," *Proceedings of the National Academy of Sciences*, 39 (10), 1095–1100.

Stubner, Ralf *dqrng: Fast Pseudo Random Number Generators* (2024) https://CRAN.R-project.org/package=dqrng, r package version 0.4.1.