A Bayesian Latent-Factor Framework for Causal Decomposition in High-Dimensional Experiments

Khaled Boughanmi, Raghuram Iyengar, and Young-Hoon Park*

Abstract

Randomized experiments with high-dimensional outcomes, such as product-level purchases, biomarker panels, or digital behavioral traces, typically yield aggregate treatment effects but provide limited insights into the underlying pathways through which interventions operate. We propose a Bayesian framework for decomposing treatment effects across latent behavioral dimensions inferred exclusively from pre-treatment data. The approach combines a causal identification strategy with a mixed-membership factor model implemented through Latent Dirichlet Allocation, yielding additive factor-level outcomes that accommodate sparsity, overlapping item–factor relationships, and uncertainty in latent assignments. We derive a causal decomposition that represents factor-level treatment effects as probability-weighted averages of item-level effects, along with an adjustment term capturing the alignment between latent structure and heterogeneous responses. A joint posterior computation scheme integrates latent factor estimation with treatment-effect inference and propagates uncertainty to all causal estimands. Simulation studies demonstrate that the method reliably recovers both latent factor structure and treatment-effect decompositions under realistic sparsity. In an application to a large-scale randomized promotion experiment at a retailer, the framework identifies interpretable latent behavioral factors, isolates the components through which the promotion operates, and reveals heterogeneity obscured by category-based analyses. The proposed method provides a scalable and causally principled approach for analyzing high-dimensional experimental outcomes.

Keywords: Causal inference; Bayesian inference; Latent factor models; Mixed-membership models; High-dimensional data; Heterogeneous treatment effects; Randomized experiments

*Khaled Boughanmi is an Assistant Professor of Marketing at the Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, NY, kb746@cornell.edu. He is the corresponding author. Raghuram Iyengar is the Miers-Busch W'1885 Professor and Professor of Marketing, Department of Marketing, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, riyengar@wharton.upenn.edu.Young-Hoon Park is the Sung-Whan Suh Professor of Management and Professor of Marketing, Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853, yp34@cornell.edu.

1 Introduction

Modern experiments in biostatistics, economics, marketing, and the social sciences increasingly generate high-dimensional outcome spaces. Consumers purchase from assortments containing thousands of products; patients present panels of biomarkers spanning multiple physiological systems; and digital platforms capture detailed behavioral traces at fine resolution. Although the average treatment effect associated with an intervention provides a stable and policy-relevant summary, it offers limited insight into how the intervention operates across the underlying behavioral, biological, or functional dimensions that shape the observed outcome. Such insights are often critical for intervention design and targeting (e.g., Imbens & Rubin 2015, Bauer et al. 2021, Zhang & Tchetgen Tchetgen 2021).

An important consideration in evaluating treatment-effect heterogeneity, and its relevance for downstream decisions such as policy design, is the choice of the dimensions along which heterogeneity is characterized. While observable features such as individual demographics are a natural starting point, relying exclusively on observables can be limiting (e.g., Rossi et al. 1996). Administrative taxonomies, such as product categories and ICD codes, often reflect managerial or historical conventions and may not align with underlying response patterns. Consequently, analyses can benefit from incorporating both observable and latent sources of heterogeneity. Conversely, avoiding any aggregation can yield unstable estimates due to data sparsity. As a result, the informational value of high-dimensional randomized experiments can be substantially increased by statistical methods that recover latent dimensions relevant for heterogeneity while preserving causal validity, and propagate uncertainty in the inferred latent structure through to final causal estimates.

Developing methods that leverage latent structure for causal inference introduces a distinct set of methodological challenges. High-dimensional outcome vectors are typically sparse, destabilizing item-level causal estimates and complicating posterior computation. Treatment effects may vary not only with observed covariates but also with latent aspects of pre-treatment behavior, motivating decompositions that capture cross-item heterogeneity. Because the latent structure is unobserved and must be inferred, uncertainty in factor assignments should propagate into causal estimates. Finally, estimating latent structure using post-treatment outcomes risks contaminating the factor definitions themselves, thereby compromising causal interpretability—a well-known challenge in high-dimensional Bayesian factor and count-data models (e.g., Blei et al. 2003, Griffiths & Steyvers 2004, Zhu et al. 2023). In causal applications, these difficulties arise in service of identification rather than prediction, making them particularly acute.

Existing empirical strategies address only limited subsets of this problem. Estimating treatment effects separately at the item level is conceptually straightforward but empirically unstable in sparse, high-dimensional environments. Aggregation using pre-defined categories mitigates sparsity but may impose partitions misaligned with meaningful behavioral constructs. Conversely, analyses based solely on aggregate outcomes enhance robustness but obscure the mechanisms through which effects operate. Machine-learning approaches for heterogeneous treatment effects capture variation along observed covariates but ignore latent dimensions (e.g., Athey & Imbens 2016, Wager & Athey 2018, Knaus 2021) . Unsupervised learning tools such as clustering or topic models can uncover latent structure but are optimized for prediction or exploration and do not guarantee causal interpretability when

the structure is estimated using post-treatment data.

In this article, we develop a Bayesian framework for decomposing treatment effects into latent behavioral components and observed covariates in high-dimensional outcome settings. The key idea is to estimate the latent factor structure exclusively from pre-treatment data, ensuring independence from treatment assignment and preserving causal interpretability. We employ a mixed-membership latent factor model—implemented via Latent Dirichlet Allocation (Blei et al. 2003)—to flexibly represent sparse and overlapping item–factor relationships. These relationships define additive factor-level outcomes that avoid double-counting while allowing items to load probabilistically onto multiple latent components. A joint Bayesian procedure integrates latent-structure learning with treatment-effect estimation, enabling full propagation of uncertainty from factor inference to causal inference. Computationally, collapsed Gibbs sampling and modular regression updates deliver scalability in high-dimensional outcome spaces (e.g., Griffiths & Steyvers 2004, Zhu et al. 2023).

Our contributions are threefold. First, we provide a causal identification result demonstrating that factor-level effects remain valid under randomized assignment when factor memberships are inferred solely from pre-treatment outcomes, paralleling identification strategies for pathway-specific and component-specific causal effects (e.g., Bauer et al. 2021). Second, we derive an analytical decomposition expressing factor-level effects as weighted averages of item-level effects plus an adjustment that captures the alignment between latent structure and heterogeneous responses. Third, we develop a posterior computation strategy that unifies mixed-membership estimation with causal inference, and show through simulations that the proposed method reliably recovers both the latent factor structure and the treatment-effect decomposition under realistic sparsity conditions.

To illustrate the value of the proposed framework, we analyze purchase data from a randomized promotion experiment conducted at a large personal-care retailer involving close to $2,000$ items. The data exhibit substantial sparsity and high dimensionality, making them particularly well suited to our approach. Our method uncovers interpretable latent behavioral factors, identifies the components through which the promotion operates, and reveals heterogeneity that conventional category-based analyses fail to capture. Only a subset of latent factors exhibits meaningfully positive effects, and these correspond closely to distinct skincare needs identified by domain experts. The resulting decomposition offers a richer and more actionable understanding of how the intervention affects consumer behavior.

The remainder of the article is organized as follows. Section 2 presents the causal framework for factor-level decomposition and describes the Bayesian latent factor model and associated inference procedure. Section 3 evaluates finite-sample performance across a range of simulation settings. Section 4 applies the approach to the randomized field experiment. Finally, Section 5 provides a brief discussion and conclusion.

2 Model and Inference

We consider an experiment with I units (e.g., consumers) indexed by $i = 1, \ldots, I$, each measured on up to J items indexed by $j = 1, \ldots, J$ (e.g., spending on individual products). Units are randomly assigned to treatment ($W_i = 1$) or control ($W_i = 0$) in a one-shot

randomized controlled trial. Let Y_{ij} denote the post-treatment outcome of unit i on item j. The unit-level aggregate outcome is $Y_i = \sum_{j=1}^{J} Y_{ij}$. Our goal is to decompose the average treatment effect (ATE) on Y_i into contributions operating through a small number of latent behavioral factors that summarize the high-dimensional vector of unit-item responses. We frame this analysis within the potential-outcomes framework for causal inference (e.g., Rubin 2005, Angrist & Pischke 2008, Imbens & Rubin 2015).

Under SUTVA and consistency, randomized treatment assignment implies $W_i \perp \{Y_{ij}(1), Y_{ij}(0) : j = 1, \ldots, J\}$ (e.g., Rubin 2005, Imbens & Rubin 2015). Consequently, W_i is also independent of any measurable function of these potential outcomes, including the aggregate outcomes constructed from $\{Y_{ij}(w)\}$. This observation underlies all decomposition results that follow.

2.1 Decomposition with Observed Factors

We first consider a benchmark case in which item–factor memberships are known and deterministically fixed. Suppose unit i associates item j with a known factor k, represented by a discrete assignment variable $z_{ij} \in \{1, \ldots, K\}$ defined prior to treatment. Define factor-level outcomes

$$Y_{ik} = \sum_{j=1}^{J} \mathbb{I}\{z_{ij} = k\}\, Y_{ij}, \qquad k = 1, \ldots, K. \tag{1}$$

By construction, $Y_i = \sum_{k=1}^{K} Y_{ik}$. Let $Y_{ik}(w)$ denote the corresponding factor-level potential outcomes under treatment $w \in \{0, 1\}$. All expectations below are taken over units.

Proposition 1 (Observed-Factor Decomposition)**.** *If treatment is randomized such that $W_i \perp \{Y_{ij}(1), Y_{ij}(0)\}$, then the aggregate ATE τ decomposes additively across the factor-level ATEs τ_k: $\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = \sum_{k=1}^{K} \mathbb{E}[Y_{ik}(1) - Y_{ik}(0)] = \sum_{k=1}^{K} \tau_k$. Moreover, the difference-in-means estimator $\hat{\tau}_k = \mathbb{E}[Y_{ik} \mid W_i = 1] - \mathbb{E}[Y_{ik} \mid W_i = 0]$ is unbiased for τ_k, and $\hat{\tau} = \sum_{k=1}^{K} \hat{\tau}_k$ is unbiased for τ.*

Proof. Because $Y_i(w) = \sum_{k=1}^{K} Y_{ik}(w)$, linearity of expectation yields the decomposition. Randomized treatment assignment $W_i \perp \{Y_{ij}(1), Y_{ij}(0)\}$, and since each $Y_{ik}(w)$ is a measurable function of the item-level potential outcomes, it follows that $W_i \perp \{Y_{ik}(1), Y_{ik}(0)\}$. Therefore, the difference-in-means estimators are unbiased. □

This result provides the structural benchmark: when item–factor memberships are observed, the aggregate ATE decomposes exactly into the factor-level ATEs.

2.2 Latent Factors and Causal Identification

In most empirical settings, the factor assignments z_{ij} are latent and must be inferred rather than observed. Define the latent factor-level potential outcomes analogously:

$$Y_{ik}(w) = \sum_{j=1}^{J} \mathbb{I}\{z_{ij} = k\}\, Y_{ij}(w). \tag{2}$$

Let $\mathcal{D}_i^{\text{pre}}$ denote pre-treatment data used to estimate posterior membership probabilities, and suppose the factor model yields $z_{ij} \sim P(z_{ij} \mid \mathcal{D}_i^{\text{pre}})$. For these factor-level outcomes to retain causal interpretability, the information used to infer latent memberships must be independent of treatment assignment. Because $\mathcal{D}_i^{\text{pre}}$ is measured prior to treatment, randomization ensures $W_i \perp \mathcal{D}_i^{\text{pre}}$ (e.g., Stuart 2010, Imbens & Rubin 2015).

Proposition 2 (Identification with Probabilistic Assignments from Pre-Treatment Data)**.** *If treatment is randomized such that $W_i \perp \{Y_{ij}(1), Y_{ij}(0), \mathcal{D}_i^{pre}\}$, and latent assignments are drawn from $P(z_{ij} \mid \mathcal{D}_i^{pre})$, then $W_i \perp z_{i,1:J}$ and $W_i \perp \{Y_{ik}(1), Y_{ik}(0)\}$. Consequently, $\hat{\tau}_k = \mathbb{E}[Y_{ik} \mid W_i = 1] - \mathbb{E}[Y_{ik} \mid W_i = 0]$ is unbiased for τ_k.*

Proof. By construction, z_{ij} depends only on $\mathcal{D}_i^{\text{pre}}$, which is independent of W_i under randomization, implying $W_i \perp z_{i,1:J}$. The factor-level potential outcomes are measurable functions of $\{Y_{ij}(w), z_{ij}\}$, both independent of W_i, so $W_i \perp \{Y_{ik}(1), Y_{ik}(0)\}$. Unbiasedness of the difference-in-means estimator then follows. □

Randomized treatment assignment thus guarantees identification of factor-level causal effects even when factor memberships are latent and probabilistically inferred. Importantly, although not strictly necessary, it is sufficient for causal interpretability that latent factor allocations be inferred exclusively from pre-treatment data.

2.3 Relation Between Item- and Factor-Level Treatment Effects

Having established identification of factor-level causal effects under both observed and latent factor memberships, we now clarify how these effects relate to the underlying unit- and item-level treatment effects $\tau_{ij} = Y_{ij}(1) - Y_{ij}(0)$ and their averages across units, $\tau_j = \mathbb{E}[\tau_{ij}]$. Recall that factor-level potential outcomes are defined by $Y_{ik}(w) = \sum_{j=1}^{J} \mathbb{I}\{z_{ij} = k\}\, Y_{ij}(w)$, so that $Y_i(w) = \sum_{k=1}^{K} Y_{ik}(w)$ for $w \in \{0, 1\}$.

Proposition 3 (Factor-Level Effects as Probability-Weighted Aggregates Plus Covariance Adjustments)**.** *Under randomized treatment assignment and factor memberships represented by latent indicators z_{ij}, $\tau_k = \mathbb{E}[Y_{ik}(1) - Y_{ik}(0)] = \sum_{j=1}^{J} \omega_{jk}\, \tau_j + \sum_{j=1}^{J} \operatorname{cov}\Big(\mathbb{I}\{z_{ij} = k\},\, \tau_{ij}\Big)$, where $\omega_{jk} = \mathbb{E}\Big[\mathbb{I}\{z_{ij} = k\}\Big]$ is the marginal probability that item j belongs to factor k.*

Proof. By definition, $\tau_k = \mathbb{E}\Big[\sum_{j=1}^{J} \mathbb{I}\{z_{ij} = k\}\tau_{ij}\Big] = \sum_{j=1}^{J} \mathbb{E}[\mathbb{I}\{z_{ij} = k\}\tau_{ij}]$. The covariance identity gives $\mathbb{E}[\mathbb{I}\{z_{ij} = k\}\, \tau_{ij}] = \omega_{jk}\, \tau_j + \operatorname{cov}\Big(\mathbb{I}\{z_{ij} = k\},\, \tau_{ij}\Big)$, and summing over j yields the result. □

This proposition shows that a factor-level effect is not merely a simple aggregation of item-level treatment effects. The first term, $\sum_{j=1}^{J} \omega_{jk}\, \tau_j$, is the probability-weighted average of item-level ATEs, where the weights reflect how likely each item belongs to factor k. The second term, $\sum_{j=1}^{J} \operatorname{cov}\Big(\mathbb{I}\{z_{ij} = k\}, \tau_{ij}\Big)$, captures how factor memberships and unit–item treatment effects co-vary across units and items.

Corollary 1 (Pure Averaging Case)**.** *If $\operatorname{cov}\Big(\mathbb{I}\{z_{ij} = k\},\, \tau_{ij}\Big) = 0$ for all j, $\tau_k = \sum_{j=1}^{J} \omega_{jk}\, \tau_j$.*

Two familiar special cases follow. First, when factor memberships do not vary across units ($z_{ij} = z_j$ for all i), the covariance term vanishes, and the factor-level effects reduce to a probability-weighted average of the item-level effects. Second, when item-level treatment effects do not vary across units ($\tau_{ij} = \tau_j$), the covariance is zero, and τ_k collapses to a probability-weighted sum of $\{\tau_j\}$.

Corollary 2 (Heterogeneity-Based Adjustments Cancel in Aggregate Effects)**.** *Because the latent indicators exhaust all factors,* $\sum_{k=1}^{K} \mathbb{I}\{z_{ij} = k\} = 1$ *for every* (i, j)*. Thus,* $\tau = \sum_{k=1}^{K} \tau_k$*, and an analogous aggregation holds at the item level:* $\tau = \sum_{j=1}^{J} \tau_j$*. In particular, although the covariance adjustments in Proposition 3 may be nonzero for individual factors, they cancel in aggregate and do not affect* τ*.*

These results highlight how factor-level causal effects relate to the underlying item-level effects. Factor-level effects remain causally identified under randomized assignment and pre-treatment factor inference, but their interpretation depends on how the factor model groups items and how those groupings align with heterogeneous treatment responses, much as in pathway- or component-specific decompositions of causal effects (e.g., Bauer et al. 2021, Zhang & Tchetgen Tchetgen 2021). An illustrative example demonstrating the decomposition is provided in Supplementary Material S1.

2.4 Latent Factor Model

To operationalize the decomposition results above, we employ a probabilistic model that infers latent factors from observed data in a manner consistent with the identification conditions in Proposition 2.

Although the causal decomposition is valid for any factor model whose assignments depend exclusively on pre-treatment information, not all models are equally suitable for empirical implementation. A practical model must satisfy several requirements. First, it should yield interpretable behavioral factors that generate meaningful insights. Second, it should infer factor memberships directly from data rather than impose restrictive or ad hoc groupings. Third, it should produce posterior item–factor probabilities that map cleanly into the assignments z_{ij} used in the causal decomposition. Finally, it should support flexible probabilistic memberships, allowing items to load on multiple factors rather than forcing mutually exclusive categories.

We adopt the Latent Dirichlet Allocation (LDA) model (Blei et al. 2003), which meets these requirements and has been widely used for high-dimensional count data in both methodological and applied work (e.g., Griffiths & Steyvers 2004, Wallach et al. 2009, Tirunillai & Tellis 2014, Zhong & Schweidel 2020). LDA treats each unit-item instance or interaction (e.g., purchase or review) as an expression of an underlying behavioral factor (e.g., topic, motive, or theme), learning both which factors drive each unit's behavior and how strongly each item associates with each factor. The model provides a transparent probabilistic structure and supports uncertainty quantification through posterior sampling methods such as Gibbs sampling (e.g., Griffiths & Steyvers 2004). LDA assumes a fixed number of latent factors that are common across units and stable across treatment conditions. It also employs a bag-of-items representation that treats item order as irrelevant—an exchangeability assumption applied within and across units.

Each latent factor k is associated with a J-dimensional factor–item distribution $\boldsymbol{\phi}_k$, where ϕ_{kj} measures the strength of association between factor k and item j. Each unit i is characterized by a K-dimensional factor-membership distribution $\boldsymbol{\theta}_i$, where θ_{ik} captures the prevalence of factor k for that unit. For each observed instance n of unit i with item v_{in}, the generative process is

$$\begin{cases} P(\eta_{in} = k \mid \boldsymbol{\theta}_i) & = \theta_{ik}, \\ P(v_{in} = j \mid \boldsymbol{\phi}_{\eta_{in}}) & = \phi_{\eta_{in} j}, \end{cases}$$

where η_{in} is the latent factor assigned to instance n and v_{in} is the observed item. Given this structure, the posterior probability that item j for unit i belongs to factor k is

$$P(z_{ij} = k \mid \boldsymbol{\theta}_i, \boldsymbol{\phi}_k) = P(\eta_{in} = k \mid v_{in} = j, \boldsymbol{\theta}_i, \boldsymbol{\phi}_k) \propto \theta_{ik}\phi_{kj}. \tag{3}$$

A key advantage of this probabilistic formulation is that it explicitly maintains uncertainty in how items relate to latent factors. When items reflect multiple behavioral processes, probabilistic rather than deterministic assignments avoid overstating the precision of factor memberships and ensure that uncertainty is coherently propagated to the factor-level treatment effect estimates τ_k. Although η_{in} (instance-level assignments) and z_{ij} (unit-item factor labels) are distinct latent variables, the conditional independence structure of LDA implies that they share the same posterior form in (3), yielding a coherent measure of item-factor association. Finally, because all parameters $(\boldsymbol{\theta}_i, \boldsymbol{\phi}_k)$—and therefore all posterior memberships $P(z_{ij} = k \mid \boldsymbol{\theta}_i, \boldsymbol{\phi}_k)$—are estimated exclusively from pre-treatment data $\mathcal{D}_i^{\text{pre}}$, the resulting factor assignments remain independent of treatment under randomized assignment. Hence, the causal interpretation of the factor-level decomposition is preserved, and LDA-based factor-level estimates remain unbiased and scientifically interpretable under the experimental design.

2.5 Estimation

We now turn to estimation of the treatment effects defined above. At the aggregate level, a common starting point is

$$Y_i = \alpha + \tau W_i + \varepsilon_i, \tag{4}$$

where Y_i is the aggregate outcome for unit i, W_i is the treatment indicator, and τ is the ATE. Under randomized assignment, the difference-in-means estimator or the OLS estimator in (4) consistently estimates τ (e.g., Angrist & Pischke 2008, Imbens & Rubin 2015).

2.5.1 Factor-Level Treatment Effects

When factor memberships are observed or inferred from pre-treatment data, factor-level outcomes Y_{ik} can be regressed analogously:

$$Y_{ik} = \alpha_k + \tau_k W_i + \varepsilon_{ik}, \tag{5}$$

where τ_k is the factor-specific ATE. Because aggregate outcomes satisfy $Y_i = \sum_{k=1}^{K} Y_{ik}$, Proposition 1 implies that the aggregate effect decomposes as $\tau = \sum_{k=1}^{K} \tau_k$. This identity is also clear from the regression structure. Summing (5) over k gives $Y_i = \sum_{k=1}^{K} \alpha_k +$

$\left(\sum_{k=1}^{K} \tau_k\right) W_i + \sum_{k=1}^{K} \varepsilon_{ik}$. Letting $\tilde{\alpha} = \sum_{k=1}^{K} \alpha_k$, $\tilde{\tau} = \sum_{k=1}^{K} \tau_k$, and $\tilde{\varepsilon}_i = \sum_{k=1}^{K} \varepsilon_{ik}$ yields $Y_i = \tilde{\alpha} + \tilde{\tau} W_i + \tilde{\varepsilon}_i$, so $\tilde{\tau} = \tau$ in the population. The decomposition $\tau = \sum_{k=1}^{K} \tau_k$ holds for any sample size. Under standard regularity conditions, estimation satisfies this linear constraint, ensuring coherent inference across both aggregate and factor levels.

2.5.2 Heterogeneity in Treatment Effects

The framework also accommodates heterogeneity in treatment effects arising from both observed and latent pre-treatment characteristics. Let $\boldsymbol{x}_i$ denote a vector of M observed pre-treatment covariates for unit i and $\boldsymbol{d}_i$ a vector of S latent features inferred exclusively from pre-treatment data. These features may include summaries of pre-treatment outcomes across factors or other transformations that capture underlying behavioral variation.

To be concrete, we set $S = K$ in our setting and construct $\boldsymbol{d}_i$ as a K-dimensional vector of total pre-treatment outcomes for each factor k, with entries

$$d_{ik} = \sum_{j=1}^{J} \mathbb{I}_k(z_{ij})\, Y_{ij}^{pre}, \tag{6}$$

where Y_{ij}^{pre} is the pre-treatment outcome for unit i on item j. This mirrors the factor-level decomposition used for post-treatment outcomes and provides interpretable, factor-specific summaries of pre-treatment behavior.

Let $\tilde{\boldsymbol{x}}_i = (\boldsymbol{x}_i, \boldsymbol{d}_i)$ denote the combined feature vector. We model the heterogeneous treatment effect (HTE) for unit i as

$$\tau_i = \boldsymbol{\tau}^{h\top}(1, \tilde{\boldsymbol{x}}_i) = \tau_0^h + \sum_{m=1}^{M} \tau_m^h \tilde{x}_{im} + \sum_{s=1}^{S} \tau_{M+s}^h \tilde{x}_{i,M+s},$$

where τ_0^h is the baseline treatment effect. Substituting τ_i into the potential-outcome representation yields

$$Y_i = \alpha_i + \tau_i W_i + \varepsilon_i^h, \tag{7}$$

where the heterogeneous baseline level is $\alpha_i = \boldsymbol{\alpha}^{h\top}(1, \tilde{\boldsymbol{x}}_i)$ and $\varepsilon_i^h \sim \mathcal{N}(0, \sigma^{2h})$. The coefficients $\boldsymbol{\tau}^h$ characterize how treatment effects vary with both observed and latent pre-treatment features, consistent with parametric HTE specifications commonly used in applied work (e.g., Angrist & Pischke 2008, Knaus 2021).

Factor-level heterogeneous effects are defined analogously:

$$\tau_{ik} = \boldsymbol{\tau}_k^{h\top}(1, \tilde{\boldsymbol{x}}_i) = \tau_{0k}^h + \sum_{m=1}^{M} \tau_{mk}^h \tilde{x}_{im} + \sum_{s=1}^{S} \tau_{M+s,k}^h \tilde{x}_{i,M+s},$$

with corresponding factor-level outcomes

$$Y_{ik} = \alpha_{ik} + \tau_{ik} W_i + \varepsilon_{ik}^h. \tag{8}$$

Because $Y_i = \sum_{k=1}^{K} Y_{ik}$ and $\tau_i = \sum_{k=1}^{K} \tau_{ik}$, the heterogeneous treatment effects satisfy the

additive decomposition $\boldsymbol{\tau}^h = \sum_{k=1}^{K} \boldsymbol{\tau}_k^h$. Table 1 summarizes the aggregate and factor-level treatment-effect representations.

Table 1: Treatment Effects at Aggregate and Factor Levels

	Outcomes	Average Treatment Effects	Heterogeneous Treatment Effects
Aggregate	Y_i	τ (Eq. 4)	$\boldsymbol{\tau}^h$ (Eq. 7)
Factor-level	Y_{ik}	τ_k (Eq. 5)	$\boldsymbol{\tau}_k^h$ (Eq. 8)

2.6 Inference

We estimate latent factors using pre-treatment data, preserving the causal interpretation of factor-level ATEs (Proposition 2). Treatment effects at both the aggregate and factor levels are then estimated while fully propagating uncertainty from the factor model.

Our inference procedure accounts for three sources of uncertainty that are essential for valid factor-level inference: unit–factor prevalences $\boldsymbol{\theta}_i$, factor–item distributions $\boldsymbol{\phi}_k$, and item–to–factor allocations z_{ij}. Rather than conditioning on point assignments, we integrate over their posterior distributions so that uncertainty in factor learning coherently propagates to τ, $\{\tau_k\}$, and heterogeneous effects.

2.6.1 Algorithm

We adopt LDA with symmetric Dirichlet priors and employ a collapsed Gibbs sampler for instance-level assignments, using only pre-treatment data to estimate $(\boldsymbol{\theta}_i, \boldsymbol{\phi}_k)$.

1. **Pre-treatment factor learning.** Specify symmetric Dirichlet priors

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\zeta, \ldots, \zeta), \qquad \boldsymbol{\phi}_k \sim \text{Dirichlet}(\beta, \ldots, \beta).$$

   Update instance-level assignments η_{in} via the standard collapsed Gibbs step (e.g., Griffiths & Steyvers 2004). Then draw

$$\boldsymbol{\theta}_i \mid \boldsymbol{n}_i^u \sim \text{Dirichlet}(\zeta + \boldsymbol{n}_i^u), \qquad \boldsymbol{\phi}_k \mid \boldsymbol{n}_k^f \sim \text{Dirichlet}(\beta + \boldsymbol{n}_k^f), \tag{9}$$

   where $\boldsymbol{n}_i^u$ and $\boldsymbol{n}_k^f$ denote the unit-level and factor-level item counts by the collapsed sampler.

2. **Item–factor association.** For each unit–item pair (i, j), sample z_{ij} from $P(z_{ij} = k \mid \boldsymbol{\theta}_i, \boldsymbol{\phi}_k) \propto \theta_{ik}\, \phi_{kj}$. Construct factor-level outcomes as $Y_{ik} = \sum_{j=1}^{J} \mathbb{I}\{z_{ij} = k\} Y_{ij}$.

3. **Heterogeneity features.** Construct latent pre-treatment features $\boldsymbol{d}_i$ by aggregating the pre-treatment outcomes to latent factors: $d_{ik} = \sum_{j=1}^{J} \mathbb{I}\{z_{ij} = k\} Y_{ij}^{\text{pre}}$. These features enter the heterogeneous-effects model.

4. **Causal estimation.** Within each MCMC iteration, estimate the aggregate and factor-level outcome regressions using Y_i, Y_{ik}, and heterogeneity features. Sampling regression parameters at every iteration propagates uncertainty in $(\boldsymbol{\theta}_i, \boldsymbol{\phi}_k, z_{ij})$ directly into posterior draws for τ, $\{\tau_k\}$, and heterogeneous effects.

Full conditionals, priors, and diagnostics are provided in Supplementary Material S2.

3 Simulations

We conduct a series of Monte Carlo simulations to evaluate the performance of the proposed framework in recovering latent factor structures and estimating treatment effects. The simulations address three questions: whether the method yields valid causal estimates under both null and non-null treatment effects; whether it accurately recovers heterogeneous treatment effects across latent factors; and whether it scales reliably to large item spaces comparable to our empirical application.

3.1 Simulation 1: Recovery of Average Treatment Effects

We begin with a controlled setting that allows transparent comparison between true and estimated factor structures. We simulate $I = 100$ units, $J = 10$ items, and $K = 3$ latent factors. Unit–factor prevalence are drawn from a symmetric Dirichlet distribution with $\zeta = 0.1$, and factor–item distributions are drawn from a symmetric Dirichlet distribution with $\beta = 0.1$, following standard mixed-membership and topic-modeling formulations (e.g., Blei et al. 2003, Griffiths & Steyvers 2004). For each unit, we generate ten unit-item instances as pre-treatment data. We randomly assign half of the units to treatment condition ($W_i = 1$) and the remainder to control ($W_i = 0$). Post-treatment item-level outcomes are then generated for each unit.

To assess whether the method artifacts treatment effects, we simulate both null and non-null treatment effect settings. For each unit–item pair (i, j) we generate outcomes according to $Y_{ij} = \tau_j W_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$, independently across all i and j. In the null case, $\tau_j = 0$ for all j. In the non-null case, item-level treatment effects are drawn independently as $\tau_j \sim \mathcal{N}(0.5, 0.25)$, $j = 1, \ldots, J$.

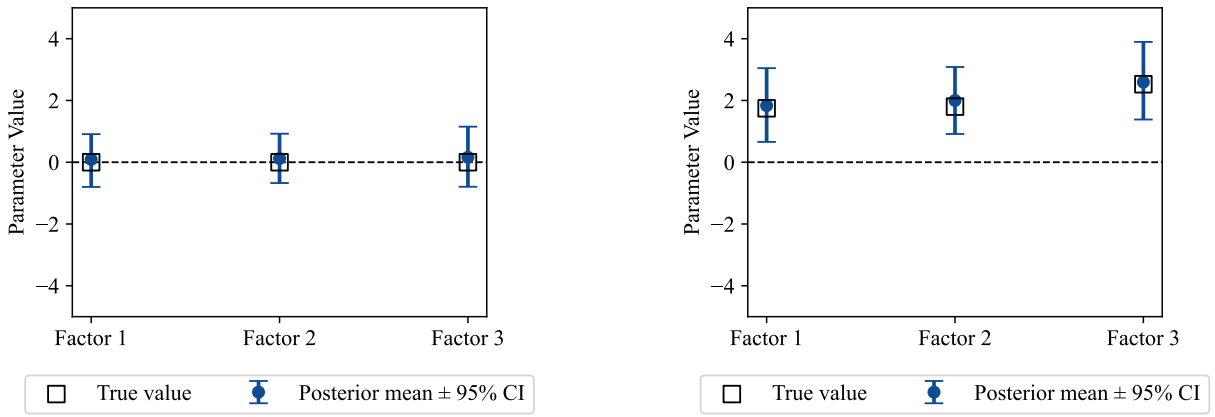
Given unit–factor weights θ_{ik} and factor–item loadings ϕ_{kj}, the item–factor assignment weight is proportional to $\theta_{ik}\,\phi_{kj}$. Because the item-level effects τ_j are constant across units by construction in this simulation, the latent factor treatment effect τ_k is a probability-weighted mixture of the underlying item-level effects, as characterized in Proposition 3. In particular, if $\tau_j = 0$ for all items, then all factor-level treatment effects are necessarily zero. Estimation follows the inference procedure described in Section 2.6. We run a Gibbs sampler for 2,000 iterations, discarding the first 1,000 as burn-in and retaining the remaining draws for posterior inference (e.g., Griffiths & Steyvers 2004, Heinrich 2005).

To assess recovery of the latent structure, we compare the true factor–item probability matrix $\Phi = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3)^\top$ with its posterior mean estimate:

$$\Phi = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.40 & 0.02 & 0.40 & 0.01 & 0.00 & 0.00 & 0.17 \\ 0.07 & 0.01 & 0.01 & 0.00 & 0.73 & 0.18 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.24 & 0.00 & 0.04 & 0.00 & 0.58 & 0.05 & 0.03 & 0.00 & 0.06 & 0.00 \end{bmatrix}$$

and

$$\widehat{\Phi} = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.42 & 0.02 & 0.38 & 0.00 & 0.00 & 0.00 & 0.19 \\ 0.03 & 0.00 & 0.01 & 0.00 & 0.73 & 0.22 & 0.00 & 0.00 & 0.00 & 0.02 \\ 0.23 & 0.00 & 0.04 & 0.00 & 0.58 & 0.05 & 0.05 & 0.00 & 0.06 & 0.00 \end{bmatrix}.$$



(a) Null Treatment Effects

(b) Non-Null Treatment Effects

Figure 1: Posterior Estimates of Factor-Level Treatment Effects Under Null and Non-Null Item-Level Effects

The estimated factor–item distributions exhibit excellent recovery, with posterior means closely aligned with the true probabilities. Figure 1 reports posterior means and 95% credible intervals for factor-level treatment effects τ_k in both the null and non-null settings. In the null case (Figure 1a), the posterior means remain tightly centered around zero after burn-in, and all 95% credible intervals include zero for every k, indicating that the method remains unbiased and does not produce spurious effects. In the non-null case (Figure 1b), the posterior means closely match the true factor-level treatment effects, and all 95% credible intervals cover the corresponding true values.

3.2 Simulation 2: Recovery of Heterogeneous Treatment Effects

Having verified that the model does not generate spurious treatment effects, we next evaluate the model's ability to recover heterogeneous treatment effects when such heterogeneity is present (e.g., Athey & Imbens 2016, Wager & Athey 2018).

We retain the same basic structure as in Simulation 1 ($I = 100$, $J = 10$, $K = 3$, $\zeta = \beta = 0.1$), but introduce heterogeneity through unit-level covariates. For each unit i, we draw $M = 4$ pre-treatment covariates $x_{im} \sim \mathcal{N}(0, 1)$, $m = 1, \ldots, M$, and collect them in the vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iM})$. The heterogeneous item-level treatment effect for unit i and item j is specified as $\tau_{ij} = \boldsymbol{\tau}_j^{h\top}(1, \boldsymbol{x}_i) = \tau_{0j}^h + \sum_{m=1}^{M} \tau_{mj}^h \, x_{im}$, where the heterogeneous loading vectors are drawn independently as $\boldsymbol{\tau}_j^h \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $j = 1, \ldots, J$. Post-treatment outcomes are then generated as $Y_{ij} = \tau_{ij} \, W_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$.

We estimate the model using the same inference procedure and MCMC settings as in Simulation 1. Figure 2 summarizes posterior estimates for heterogeneous effects at the factor level. The recovered heterogeneous coefficients closely match the true data-generating loadings, and credible intervals concentrate around the true values, indicating that the framework successfully captures covariate-driven heterogeneity across latent factors.

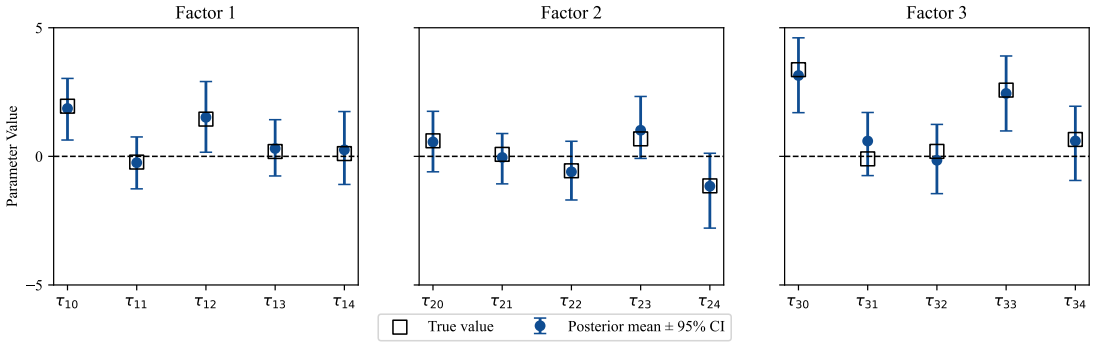


Figure 2: Posterior Estimates of Heterogeneous Factor-Level Treatment Effects

3.3 Simulation 3: Scalability to Large Item Spaces

Finally, we assess scalability in a setting aligned with our empirical application and with high-dimensional factor and count-data models more broadly (e.g., Mimno et al. 2012, Zhu et al. 2023). We simulate $I = 2,000$ units and $J = 1,000$ items, with each unit generating 30 pre-treatment observations. Latent factors are generated analogously to Simulation 1 using the same Dirichlet hyperparameters, with $K = 10$ factors. Item-level treatment effects are drawn independently as $\tau_j \sim \mathcal{N}(0, 0.5)$, $j = 1, \ldots, J$, and post-treatment outcomes are generated as $Y_{ij} = \tau_j W_i + \varepsilon_{ij}$.

We estimate the model using the same LDA-based factor learning and regression procedure described in Section 2.6. Figure 3 reports posterior means and 95% credible intervals for the factor-level treatment effects τ_k. The posterior summaries closely track the true factor-level effects and remain tightly concentrated, indicating that the method maintains accurate recovery and numerical stability even with large item assortments. Overall, these

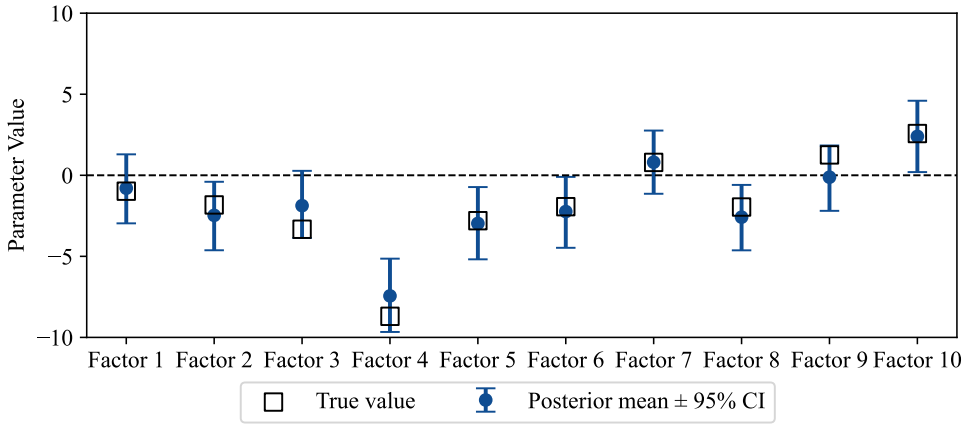


Figure 3: Posterior Estimates of Factor-level Treatment Effects in a High-Dimensional Setting ($J = 1,000$)

simulations demonstrate that the proposed framework accurately recovers latent factor compositions, preserves unbiasedness when treatment effects are null, correctly identifies and quantifies nonzero and heterogeneous effects, and scales effectively to high-dimensional outcome spaces.

4 Empirical Analysis

4.1 Experimental Data

We evaluate the proposed Bayesian decomposition of treatment effects using data from a large-scale randomized field experiment on coupon promotions originally reported in Gopalakrishnan & Park (2021). The experiment was conducted by a leading Asian personal-care retailer interested in understanding how financial incentives shape consumer purchase behavior, in line with a broader literature on field experiments in marketing (e.g., Li et al. 2015, Goldfarb et al. 2022). Consumers were randomly assigned to receive either a standard coupon ($7 off a $20 purchase), a premium coupon ($10 off a $20 purchase), or no coupon (control). Consistent with our identification results and simulation setup, we focus on the comparison between the premium-coupon and control conditions, which yields a clean experimental contrast for evaluating the inferential and interpretive value of the proposed latent-factor framework.

This empirical setting aligns well with our methodological goals. Random assignment ensures internal validity for causal effect estimation (e.g., Imbens & Rubin 2015), while the retailer's full assortment spans more than 10,000 distinct SKUs, generating an exceptionally high-dimensional and sparse item space with rich cross-product correlation patterns. The firm organizes its full assortment into five broad product categories (i.e., skincare, makeup, hair care, bath and body care, and a miscellaneous "other" category), which serve as benchmark structures for standard category-level analyses. Our objective is not to re-evaluate the managerial effectiveness of the coupon campaign, but rather to test whether latent behavioral factors inferred from pre-treatment data provide a coherent decomposition of aggregate treatment effects and sharper interpretive insight relative to these category-based benchmarks.

The analytic sample includes 4,247 consumers, of whom 3,446 were assigned to treatment and 801 to control. Within this sample, consumers purchase roughly 2,000 distinct SKUs, which define the item universe for our empirical analysis. For each consumer, we observe detailed transaction histories over a 12-month pre-treatment period and during the one-week campaign window, including product identifiers, timestamps, product attributes, and basic demographics. Randomization checks reported in Gopalakrishnan & Park (2021) confirm baseline balance across observed covariates; we build on those results by providing additional latent-balance diagnostics below. Our primary outcome is total spending (USD) during the campaign period.

4.2 Latent Factor Calibration

We estimate latent behavioral factors using an LDA model applied exclusively to pre-treatment transactions, ensuring that factor definitions remain strictly exogenous to treatment assignment and preserving the causal interpretation of factor-level treatment effects. To select the number of latent factors, we fit models with $K \in \{5, 10, \ldots, 30\}$ and evaluate predictive accuracy using perplexity, a standard out-of-sample probabilistic metric for topic models (e.g., Blei et al. 2003, Heinrich 2005). For each specification, perplexity is computed on a 10% holdout sample obtained by partitioning pre-treatment data into 90% training and 10% validation sets; lower values indicate better predictive fit.

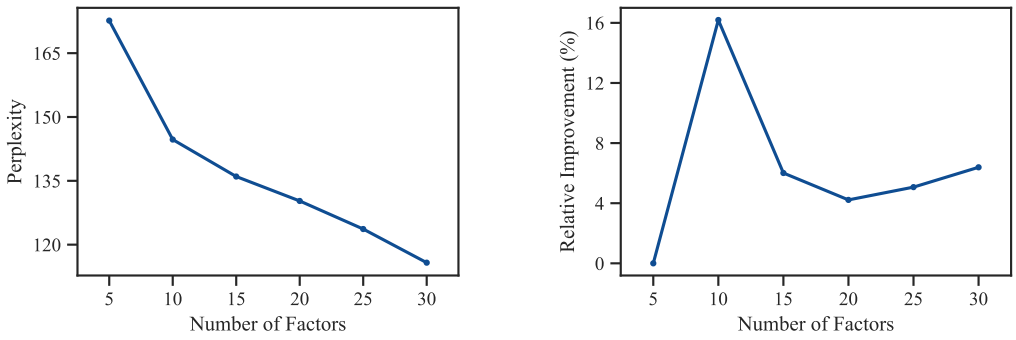


Figure 4: Holdout Perplexity and Marginal Improvements by Number of Factors

Figure 4 reports both the holdout perplexity and the marginal improvement relative to adjacent values of K. Perplexity declines monotonically with K, as expected, but marginal gains flatten considerably after $K = 10$. Balancing this empirical bias–variance trade-off with the need for interpretability, we select $K = 10$ for all subsequent analyses. The final model is estimated using 2,000 MCMC iterations. Visual inspection of the sampling chains indicates good mixing and satisfactory convergence; additional diagnostics are provided in Supplementary Material S3.

To assess the relative adequacy of the proposed mixed-membership factorization, we compare its predictive performance against three benchmarks: a degenerate membership model imposing $z_{ij} = z_j$ for all units; a model based on firm-defined product categories; and a model that assigns items randomly to five categories. The mixed-membership LDA consistently outperforms all benchmarks in out-of-sample perplexity, indicating a more faithful probabilistic representation of unit-item interaction patterns. Additional details are provided in Supplementary Material S4.

4.3 Interpretation of Latent Factors

The ten inferred latent factors provide a parsimonious yet flexible representation of pre-treatment purchase behavior, similar in spirit to other applications of topic models and mixed-membership structures in marketing (e.g., Tirunillai & Tellis 2014, Jacobs et al. 2016, Liu & Toubia 2018, Jacobs et al. 2021, Kim & Zhang 2023). Each factor captures a distinct—but not mutually exclusive—pattern of consumer behavior that extends beyond the retailer's administrative product taxonomy (e.g., skincare, makeup). Table 2 summarizes the factor structure using representative product types. Factor labels were selected in consultation with the retailer and reflect interpretable behavioral dimensions under anonymization constraints. Several factors resemble familiar product groupings (e.g., hydration-focused regimens, daily-use essentials, or gifting-oriented assortments), whereas others reveal cross-category behavioral associations not captured by the firm's predefined categories.

A key feature of the model is its probabilistic allocation of products to factors, allowing items to load on multiple behavioral dimensions (e.g., Blei et al. 2003, Airoldi et al. 2006). Figure 5 displays a t-distributed stochastic neighbor embedding (t-SNE) visualization of skincare products, colored by the modal factor assignment. The figure reveals coherent

but overlapping clusters, indicating that the factors capture multi-dimensional behavioral patterns—such as intensive hydration routines, botanical-oriented preferences, or gifting bundles—that do not map cleanly to traditional administrative categories. Approximately 52% of products exhibit non-trivial posterior mass on more than one factor (Supplementary Material S5), underscoring the value of a latent, rather than fixed, grouping structure.

Table 2: Latent Factors and Representative Product Types

	Factor Name	Representative Product Types
1	Natural Soothing Care	moisturizing mask, nourishing mask, soothing mask, mask pack, brightening mask, massage mask, hydrating mask, firming mask, air cushion, toner
2	Functional Daily Care	essence mask, fit mask, pouch pack, ampoule mask, hand mask, blackhead nose patch, cleanser, hair treatment, body lotion, soothing gel
3	Gentle & Sensitive Skincare	hydrogel mask, herb soap, toothbrush, hair treatment, toner, emulsion, vitamin supplement, balancing water, emulsion, shampoo
4	Intensive Hydration	air cushion, moisturizer, sun protector, skin refiner, foundation, essence, toner, emulsion, emulsion, sleeping mask
5	Purifying & Wellness Boost	blackhead patch, health supplement, foundation, energy ampoule, cleansing oil, cleanser, toner, scrub, slimming supplements, hair serum
6	Makeup & Daily Essentials	cushion puff, toner, lip tint, cleanser, BB cushion, eyeshadow, emulsion, sunscreen, foundation, cleansing wipes
7	Natural Indulgence & Gifting	body mist, cream, facial mask, promotional gift, cleanser, hydrating ampoule, cleansing oil, nail polish, bag & pouch, sun cushion
8	Oral Hygiene & Scalp Health	toothpaste, shampoo, mouthwash, hair treatment, shampoo, shampoo, toothpaste, shampoo, hair pack, hair treatment
9	Botanical Skincare Rituals	facial mask, toner, cream, cushion, emulsion, essence, sun protector, cleanser, serum, cleansing oil
10	Beauty Accessories & Tools	cushion puff, cotton pads, nail polish, lip & eye remover, hair mask, eyeshadow, eyelashes, nail polish, nail color, foundation

Note: Product descriptions are anonymized under data-sharing agreements. Factor labels were validated with the retailer and selected for expository purposes.

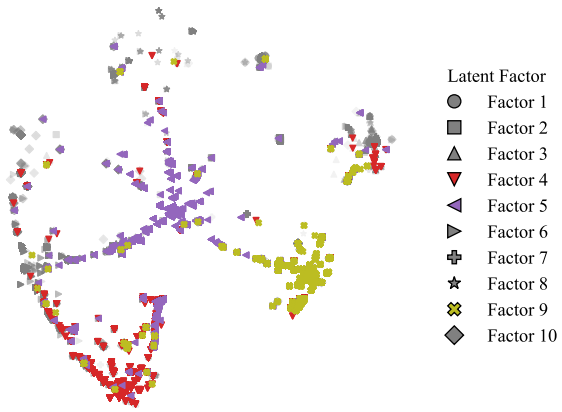


Figure 5: t-SNE Visualization of Skincare Products by Modal Factor Assignment

As a diagnostic for causal validity, Figure 6 compares pre-treatment factor distributions between treatment and control groups. Posterior mean factor shares and 95% credible intervals reveal no meaningful differences across groups for any factor. This confirms that randomization achieves balance not only on observed covariates, as documented in Gopalakrishnan & Park (2021), but also on the inferred latent structure, consistent with standard balance diagnostics in causal inference (e.g., Rosenbaum & Rubin 1983, Austin

2009). These results support the use of factor-level responses as causally valid components for treatment-effect decomposition.

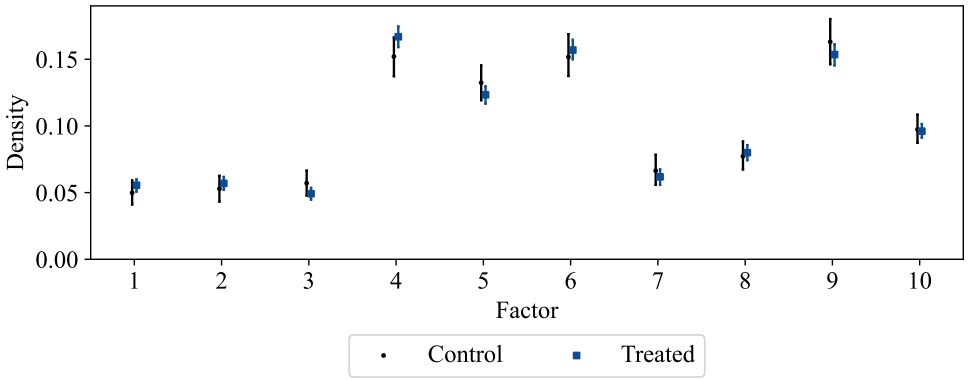


Figure 6: Pre-treatment Latent Factor Distributions by Treatment Group

4.4 Average Treatment Effects

We begin by estimating the aggregate benchmark effect. The posterior mean ATE of the coupon treatment on total campaign spending is \$1.93, with a 95% credible interval of [1.07, 2.80], consistent with the findings in Gopalakrishnan & Park (2021).

We then decompose this effect using the inferred latent behavioral factors. For each factor, we construct factor-level spending by weighting products according to their posterior probabilities of factor membership and estimate factor-specific ATEs under Bayesian framework described in Section 2. Results are reported in Table 3. Three factors exhibit credibly positive effects: Factor 9 (Botanical Skincare Rituals; \$0.64), Factor 4 (Intensive Hydration; \$0.41), and Factor 5 (Purifying & Wellness Boost; \$0.35), each with 95% credible intervals that exclude zero. The posterior means of factor-level ATEs sum to \$1.94 (95% credible interval: [1.14, 2.75]), essentially matching the aggregate effect and providing a behaviorally grounded decomposition of the overall impact.

For comparison, Table 4 reports ATEs aggregated to the retailer's five administrative product categories. The skincare category shows the largest increase (\$1.52, 95% credible interval: [0.78, 2.20]), while the remaining categories exhibit no clear effects. Relative to this coarse partition, the factor-based decomposition offers a sharper and statistically coherent interpretation. Specifically, the aggregate skincare effect is not homogeneous but instead concentrated along three behaviorally distinct latent dimensions that differ in both spending composition and effect magnitude. Thus, the latent-factor representation preserves the original experimental conclusions while extending standard difference-in-means analysis (e.g., Angrist & Pischke 2008, Imbens & Rubin 2015) by revealing interpretable structure within high-dimensional outcomes—structure that remains hidden when relying solely on administrative categories.

Table 3: Factor-Level Treatment Effects

Factor	Estimate	95% Credible Interval
1	0.08	[−0.04, 0.18]
2	0.04	[−0.07, 0.14]
3	0.08	[−0.14, 0.30]
4	0.41*	[0.02, 0.87]
5	0.35*	[0.01, 0.67]
6	0.19	[−0.05, 0.44]
7	0.06	[−0.07, 0.17]
8	0.06	[−0.09, 0.22]
9	0.64*	[0.16, 1.15]
10	0.03	[−0.12, 0.18]

Note: * indicates 95% Bayesian credible interval excludes 0.

Table 4: Category-Level Treatment Effects

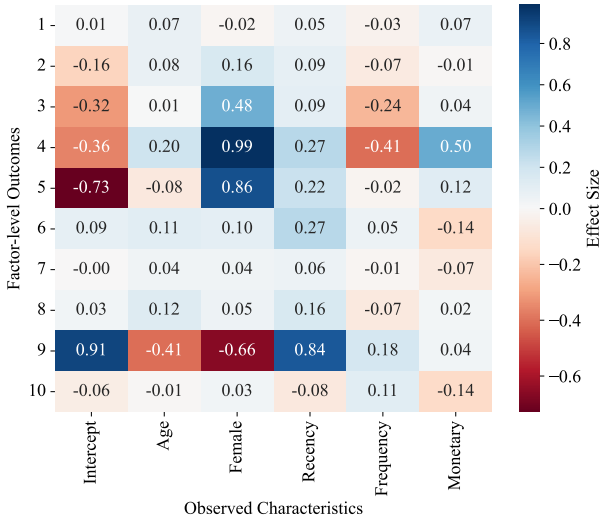
Category	Estimate	95% Credible Interval
Skin Care	1.51*	[0.07, 0.64]
Makeup	0.01	[−0.22, 0.26]
Hair Care	0.00	[−0.00, 0.00]
Bath & Body Care	0.08	[−0.14, 0.30]
Other	0.33*	[0.07, 0.64]

Note: * indicates 95% Bayesian credible interval excludes 0.

4.5 Heterogeneous Treatment Effects

We next examine whether the coupon's impact varies systematically across consumers and along latent behavioral dimensions. We consider two sources of heterogeneity: observed pre-treatment characteristics—gender, age, and RFM (recency, frequency, monetary) metrics—and unobserved heterogeneity captured by pre-treatment spending across inferred behavioral factors (e.g., Rossi et al. 1996, Kumar & Shah 2004, Fader et al. 2022). Gender is coded as a binary indicator (1 = female). The remaining characteristics are dichotomized at their median values; for recency, customers with more recent activity are coded as 1. Median-splitting the monetary variable removes its collinearity with pre-treatment factor-level spending, allowing the latter to enter the unobserved-heterogeneity analyses without multicollinearity concerns.

Figure 7 reports factor-level ATEs by subgroups defined by observed characteristics. While most subgroup contrasts exhibit overlapping credible intervals, two patterns emerge. Female customers show larger treatment effects in Factors 4 and 5, and customers with more recent activity display stronger responses in Factor 9. These patterns illustrate how latent factors uncover interpretable behavioral heterogeneity that does not align neatly with administrative product categories.

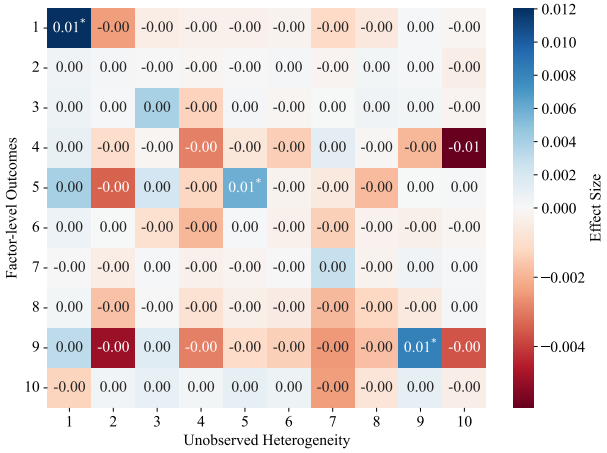


Note: * indicates 95% Bayesian credible interval excludes 0.

Figure 7: Observed Heterogeneity in Factor-Level Treatment Effects

To investigate unobserved heterogeneity, we regress factor-level treatment effects on pre-treatment spending in each corresponding latent factor, paralleling broader efforts to characterize heterogeneous responses using flexible treatment-effect models (e.g., Athey & Imbens 2016, Wager & Athey 2018). Figure 8 reports the corresponding coefficients. A clear reinforcement pattern emerges: customers with higher pre-treatment spending in a given factor tend to exhibit larger treatment effects in the same factor. The diagonal coefficients are uniformly positive, with credibly positive effects for Factors 1, 5, and 9. An additional dollar of pre-treatment spending in these factors is associated with approximately a one-cent increase in the corresponding treatment effect. This pattern suggests that the promotion coupon primarily amplifies existing behavioral tendencies rather than inducing substitution

across latent factors.



Note: * indicates 95% Bayesian credible interval excludes 0.

Figure 8: Unobserved Heterogeneity in Factor-Level Treatment Effects

For comparison, analogous analyses using the retailer's administrative product categories (Supplementary Material S6) reveal weaker and less interpretable patterns: reinforcement effects diffuse substantially and are often statistically indistinguishable from zero when outcomes are aggregated to broad categories. This contrast highlights how fixed taxonomies obscure much of the causal heterogeneity that the latent-factor approach is able to recover.

In summary, these results show that the proposed Bayesian latent-factor framework not only preserves the aggregate experimental benchmark but also yields a coherent behavioral decomposition of treatment effects and uncovers systematic heterogeneity—both observed and latent—that remains largely hidden under conventional category-based analyses. By isolating where treatment effects originate and for whom they are strongest, the framework provides a more behaviorally grounded understanding of intervention response.

5 Conclusion

This article develops a Bayesian framework for decomposing treatment effects into latent components in high-dimensional experimental settings. Motivated by modern experiments that generate sparse, multi-item outcome vectors, we show how mixed-membership factor models (e.g., Blei et al. 2003, Airoldi et al. 2006)—estimated exclusively from pre-treatment data—can be combined with randomized assignment to produce causally interpretable factor-level effects (e.g., Imbens & Rubin 2015). Restricting latent-structure learning to pre-treatment outcomes preserves the independence conditions required for causal identification while allowing items to load flexibly and probabilistically onto multiple latent factors.

Methodologically, the framework integrates three elements: a causal identification result establishing the validity of factor-level effects under randomized assignment and pre-treatment factor inference; an analytical decomposition expressing factor-level effects as probability-weighted averages of item-level effects plus an adjustment capturing alignment between latent structure and heterogeneous responses; and a joint posterior computation strategy

that propagates uncertainty in latent structure through to causal estimands. Together, these components provide a principled basis for studying how interventions operate across latent dimensions embedded in high-dimensional outcomes.

Simulation studies show that the method recovers both the latent factor structure and the induced treatment-effect decomposition under realistic sparsity conditions, including settings where item-level analyses are unstable and predefined groupings obscure meaningful heterogeneity. In an application to a large-scale randomized promotion experiment involving around 2,000 products, the framework uncovers interpretable latent behavioral factors and reveals that only a small subset of latent factors responds meaningfully to the intervention, yielding insights that conventional aggregate or category-based analyses fail to reveal.

Beyond decomposition, the framework also yields actionable diagnostics for intervention design and targeting. By estimating heterogeneous factor-level treatment effects, the model identifies the latent behavioral dimensions along which an intervention is amplified or attenuated for each unit. This enables practitioners to detect negative side effects, attribute them to specific behavioral factors, and trace those effects to the underlying items most strongly associated with each factor. These diagnostics provide a direct path to intervention design and targeting—for example, by targeting items linked to underperforming factors. An illustration of how these insights can be operationalized is provided in Supplementary S7.

Several avenues for future work remain. One direction is to treat the number of latent factors as unknown, potentially through nonparametric priors or automatic relevance determination, building on the broader literature on flexible Bayesian factor and topic models (e.g., Wallach et al. 2009, Mimno et al. 2012, Zhu et al. 2023). Another is to extend the framework to incorporate temporal evolution of latent structure while maintaining causal interpretability, enabling applications to longitudinal or sequential experiments. The approach may also be adapted to quasi-experimental settings with appropriate identification strategies (e.g., Goldfarb et al. 2022, Li & Sonnier 2023). Finally, incorporating additional structure, such as domain knowledge, hierarchical item relationships, or geometric constraints, may further enhance interpretability in applied settings.

In summary, the proposed latent-factor causal decomposition offers a general and extensible approach for analyzing high-dimensional experimental outcomes. By combining principled causal identification with flexible latent-structure learning, it provides a statistically grounded tool for uncovering the components through which interventions exert their effects.

SUPPLEMENTARY MATERIAL

The supplementary materials (PDF) provide the full conditional distributions used to estimate treatment effects at each level of analysis, along with the complete estimation algorithm. They also report convergence diagnostics for the MCMC draws in the empirical application and investigate the number of unique latent factors associated with items in the empirical setting. In addition, we describe an alternative analysis that relies on the firm's predefined product categories to decompose treatment effects and to characterize pre-treatment interactions between units and items. Finally, the supplement includes an illustrative application of the proposed approach that identifies undesirable unit–factor treatment effects and proposes actionable strategies based on the most relevant items associated with each unit–factor pair.

Funding Details

The authors have no funding details to report and no conflicts of interest to declare.

Data Availability

The data used in this study are subject to a non-disclosure agreement (NDA) with the partner firm and cannot be shared publicly.

Generative Artificial Intelligence (AI)

Generative AI tools (OpenAI's ChatGPT 5.1) were used solely for grammar refinement and non-substantive code cleaning.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., Xing, E. P. & Jaakkola, T. (2006), Mixed membership stochastic block models for relational data with application to protein-protein interactions, *in* 'Proceedings of the international biometrics society annual meeting', Vol. 15, p. 1.

Angrist, J. D. & Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion,*, Princeton University Press.

Athey, S. & Imbens, G. (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences* **113**(27), 7353–7360.

Austin, P. C. (2009), 'Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples', *Statistics in Medicine* **28**(25), 3083–3107.

Bauer, D. J., Cai, B., Hudgens, M. G. & Tchetgen Tchetgen, E. J. (2021), 'Targeted learning of pathway-specific effects', *Journal of the American Statistical Association* **116**(536), 1759–1771.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of Machine Learning Research* **3**(Jan), 993–1022.

Fader, P., Hardie, B. & Ross, M. (2022), *The Customer-Base Audit: The First Step on the Journey to Customer Centricity*, The Wharton School Press.

Goldfarb, A., Tucker, C. & Wang, Y. (2022), 'Conducting research in marketing with quasi-experiments', *Journal of Marketing* **86**(3), 1–20.

Gopalakrishnan, A. & Park, Y.-H. (2021), 'The impact of coupons on the visit-to-purchase funnel', *Marketing Science* **40**(1), 48–61.

Griffiths, T. L. & Steyvers, M. (2004), 'Finding scientific topics', *Proc. Natl. Acad. Sci. U.S.A.* **101**, 5228–5235.

Heinrich, G. (2005), Parameter estimation for text analysis, Technical report, Technical report Darmstadt, Germany.

Imbens, G. W. & Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Jacobs, B., Fok, D. & Donkers, B. (2021), 'Understanding large-scale dynamic purchase behavior', *Marketing Science* **40**(5), 844–870.

Jacobs, B. J. D., Donkers, B. & Fok, D. (2016), 'Model-based purchase predictions for large assortments', *Marketing Science* **35**(3), 389–404.

Kim, M. & Zhang, J. (2023), 'Discovering online shopping preference structures in large and frequently changing store assortments', *Journal of Marketing Research* **60**(4), 665–686.

Knaus, M. (2021), 'Heterogeneous treatment effect estimation: A review of methods and applications', *Annals of Economics and Statistics* (143), 1–42.

Kumar, V. & Shah, D. (2004), 'Building and sustaining profitable customer loyalty for the 21st century', *Journal of Retailing* **80**(4), 317–329.

Li, J. Q., Rusmevichientong, P., Simester, D. I., Tsitsiklis, J. N. & Zoumpoulis, S. I. (2015), 'The value of field experiments', *Management Science* **61**(7), 1722–1740.

Li, K. T. & Sonnier, G. P. (2023), 'Statistical inference for the factor model approach to estimate causal effects in quasi-experimental settings', *Journal of Marketing Research* **60**(3), 449–472.

Liu, J. & Toubia, O. (2018), 'A semantic approach for estimating consumer content preferences from online search queries', *Marketing Science* **37**(6), 930–952.

Mimno, D., Hoffman, M. & Blei, D. (2012), 'Sparse stochastic inference for latent dirichlet allocation', *arXiv preprint arXiv:1206.6425* .

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Rossi, P. E., McCulloch, R. E. & Allenby, G. M. (1996), 'The value of purchase history data in target marketing', *Marketing Science* **15**(4), 321–340.

Rubin, D. B. (2005), 'Causal inference using potential outcomes: Design, modeling, decisions', *Journal of the American statistical Association* **100**(469), 322–331.

Stuart, E. A. (2010), 'Matching methods for causal inference: A review and a look forward', *Statistical Science* **25**(1), 1–21.

Tirunillai, S. & Tellis, G. J. (2014), 'Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation', *Journal of Marketing Research* **51**(4), 463–479.

Wager, S. & Athey, S. (2018), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* **113**(523), 1228–1242.

Wallach, H., Mimno, D. & McCallum, A. (2009), 'Rethinking lda: Why priors matter', *Advances in neural information processing systems* **22**.

Zhang, Z. & Tchetgen Tchetgen, E. J. (2021), 'Mediation analysis under interference', *Journal of the American Statistical Association* **116**(536), 1824–1833.

Zhong, N. & Schweidel, D. A. (2020), 'Capturing changes in social media content: A multiple latent changepoint topic model', *Marketing Science* **39**(4), 827–846.

Zhu, Z., Banerjee, A. et al. (2023), 'Sparse bayesian group factor model for feature interactions in multiple count tables', *Journal of the American Statistical Association* **120**(550), 723–736.

Supplementary Material for "A Bayesian Latent-Factor Framework for Causal Decomposition in High-Dimensional Experiments"

S1 Illustrative Example of the Decomposition

This section provides a simple example that illustrates the decomposition in Proposition 3 and shows how factor-level treatment effects can differ from probability-weighted averages of item-level effects due to covariance adjustments.

S1.1 Setup

Suppose there are $I = 2$ units, $J = 3$ items, and $K = 2$ latent factors. The unit-item treatment effects $\tau_{ij} = Y_{ij}(1) - Y_{ij}(0)$ are:

	$j=1$	$j=2$	$j=3$
$i=1$	2	4	0
$i=2$	1	1	3

The corresponding item-level ATEs are:

$$\tau_1 = 1.5, \qquad \tau_2 = 2.5, \qquad \tau_3 = 1.5.$$

Latent factor memberships $z_{ij} \in \{1, 2\}$ are:

	$j=1$	$j=2$	$j=3$
$i=1$	1	1	2
$i=2$	2	2	1

Thus, each item is assigned to each factor with probability $\frac{1}{2}$:

$$\omega_{jk} = \mathbb{E}[\mathbb{I}\{z_{ij} = k\}] = \tfrac{1}{2}, \qquad \text{for all } j, k.$$

S1.2 Factor-Level Treatment Effects

By definition,

$$\tau_{ik} = \sum_{j=1}^{3} \mathbb{I}\{z_{ij} = k\}\, \tau_{ij}.$$

Computing the factor-level effects for each unit:

$$\begin{aligned} &\text{Unit 1}: \quad \tau_{11} = 2 + 4 = 6, \quad \tau_{12} = 0, \\ &\text{Unit 2}: \quad \tau_{21} = 3, \qquad\qquad \tau_{22} = 1 + 1 = 2. \end{aligned}$$

Averaging across units gives:

$$\tau_1 = \frac{6+3}{2} = 4.5, \qquad \tau_2 = \frac{0+2}{2} = 1.$$

S1.3 Decomposition Into Averages and Covariances

The probability-weighted average component is:

$$\sum_{j=1}^{3} \omega_{jk}\, \tau_j = \tfrac{1}{2}(1.5 + 2.5 + 1.5) = 2.75, \qquad k = 1, 2.$$

Next, compute the covariance terms $\text{cov}(\mathbb{I}\{z_{ij} = k\}, \tau_{ij})$:

	$j=1$	$j=2$	$j=3$
$k=1$	0.25	0.75	0.75
$k=2$	-0.25	-0.75	-0.75

Summing across items:

$$\sum_{j=1}^{3} \text{cov}(\mathbb{I}\{z_{ij} = 1\}, \tau_{ij}) = 1.75, \qquad \sum_{j=1}^{3} \text{cov}(\mathbb{I}\{z_{ij} = 2\}, \tau_{ij}) = -1.75.$$

S1.4 Verification of Proposition 3

For $k = 1$:

$$\tau_1 = 2.75 + 1.75 = 4.5.$$

For $k = 2$:

$$\tau_2 = 2.75 - 1.75 = 1.$$

These match the factor-level ATEs computed directly from potential outcomes, verifying:

$$\tau_k = \sum_{j=1}^{3} \omega_{jk}\, \tau_j + \sum_{j=1}^{3} \text{cov}\Big(\mathbb{I}\{z_{ij} = k\}, \tau_{ij}\Big).$$

Finally, the aggregate ATE is:

$$\tau = \sum_{j=1}^{3} \tau_j = 5.5 = \sum_{k=1}^{2} \tau_k,$$

demonstrating Corollary 2: covariance adjustments may distort factor-level effects but always cancel in aggregate.

S2 Full Conditionals for Regression Models

This section provides the full conditional distributions for the regression models used in the main text. All models take the generic form

$$Y_i^{\text{target}} = \boldsymbol{\tau}^{\text{target}\top} \boldsymbol{\chi}_i^{\text{target}} + \varepsilon_i^{\text{target}}, \qquad \varepsilon_i^{\text{target}} \sim \mathcal{N}(0, \sigma^{2\,\text{target}}), \tag{S2.1}$$

where "target" indexes the aggregate ATE, factor-level ATEs, and the heterogeneous-effects specifications

Table S2.1: Aggregate and Factor-Level Regression Models

Model	Y_i^{target}	$\boldsymbol{\chi}_i^{\text{target}}$	$\boldsymbol{\tau}^{\text{target}}$	$\sigma^{2\text{target}}$
Average Treatment Effects				
Aggregate	Y_i	$(1, W_i)$	$\boldsymbol{\tau} = (\alpha, \tau)$	σ^2
Factor-Level	Y_{ik}	$(1, W_i)$	$\boldsymbol{\tau}_k = (\alpha_k, \tau_k)$	σ_k^2
Heterogeneous Treatment Effects				
Aggregate	Y_i	$(1, \tilde{\boldsymbol{x}}_i, W_i, \tilde{\boldsymbol{x}}_i \cdot W_i)$	$\boldsymbol{\tau}^h = (\boldsymbol{\alpha}^h, \boldsymbol{\tau}^h)$	σ^{2h}
Factor-Level	Y_{ik}	$(1, \tilde{\boldsymbol{x}}_i, W_i, \tilde{\boldsymbol{x}}_i \cdot W_i)$	$\boldsymbol{\tau}_k^h = (\boldsymbol{\alpha}_k^h, \boldsymbol{\tau}_k^h)$	σ_k^{2h}

S2.1 Priors

We use conjugate Normal–Inverse-Gamma priors: $\boldsymbol{\tau}^{\text{target}} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\sigma^{2\,\text{target}} \sim \text{Inv-Gamma}(a_0, b_0)$, where a_0 and b_0 are shape and scale parameters. These priors apply to all regression specifications in Table S2.1.

S2.2 Full Conditionals

Let I denote the number of units in the target regression, $\boldsymbol{y}^{\text{target}} = (Y_i^{\text{target}})_{i=1}^I$, and $\boldsymbol{\chi}^{\text{target}}$ the $I \times p$ design matrix.

(1) Regression coefficients.

$$\boldsymbol{\tau}^{\text{target}} \mid \boldsymbol{y}^{\text{target}}, \boldsymbol{\chi}^{\text{target}}, \sigma^{2\,\text{target}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{S2.2}$$

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^{2\,\text{target}}} \boldsymbol{\chi}^{\text{target}\top} \boldsymbol{\chi}^{\text{target}} \right)^{-1}, \tag{S2.3}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^{2\,\text{target}}} \boldsymbol{\chi}^{\text{target}\top} \boldsymbol{y}^{\text{target}} \right). \tag{S2.4}$$

(2) Error variance.

$$\sigma^{2\,\text{target}} \mid \boldsymbol{y}^{\text{target}}, \boldsymbol{\chi}^{\text{target}}, \boldsymbol{\tau}^{\text{target}} \sim \text{Inv-Gamma}(a, b), \tag{S2.5}$$

$$a = a_0 + \frac{I}{2}, \tag{S2.6}$$

$$b = b_0 + \frac{1}{2}\left(\boldsymbol{y}^{\text{target}} - \boldsymbol{\chi}^{\text{target}}\boldsymbol{\tau}^{\text{target}}\right)^{\top}\left(\boldsymbol{y}^{\text{target}} - \boldsymbol{\chi}^{\text{target}}\boldsymbol{\tau}^{\text{target}}\right). \tag{S2.7}$$

These full conditionals yield closed-form Gibbs updates in each MCMC iteration.

(3) Pre-treatment Factor Assignments. For each pre-treatment instance n of unit i, we sample the factor label η_{in} via the collapsed Gibbs update:

$$P(\eta_{in} = k | \eta_{\neg in}, \boldsymbol{v}, \zeta, \beta) \propto (n^{u}_{ik,\neg in} + \zeta) \cdot \frac{(n^{f}_{kv_{in},\neg in} + \beta)}{\sum_{j=1}^{J} n^{f}_{kj,\neg in} + J\beta}, \tag{S2.8}$$

where v_{in} is the observed item, $n^{u}_{ik,\neg in}$ is the number of unit-factor assignments excluding instance n of unit i, and $n^{f}_{kj,\neg in}$ is the number of item-factor assignments excluding instance n of unit i. See Griffiths & Steyvers (2004) for details.

S2.3 Hyperparameters

We use weakly informative priors: $\zeta = 0.1$, $\beta = 0.01$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = 10^2\,\mathbf{I}$, and $a_0 = b_0 = 1$. These settings impose minimal prior structure while maintaining numerical stability in the Gibbs sampler. They also encourage units to load on a limited number of factors while allowing each factor to maintain diffuse support across items.

Algorithm S2.1 summarizes the full inference procedure for estimating latent factors and treatment effects while propagating factor-model uncertainty.

Algorithm S2.1 Joint Estimation of Latent Factors and Treatment Effects via Collapsed Gibbs Sampling

Initialization
Randomly assign each pre-treatment instance η_{in} to a factor $k \in \{1, \ldots, K\}$.
Compute initial collapsed Gibbs counts:
  n^u_{ik}: number of pre-treatment instances of unit i assigned to factor k
  n^f_{kj}: number of times item j is assigned to factor k across all units
for MCMC iteration $t = 1, \ldots, T$ **do**
  (A) Collapsed Gibbs Sampling for Instance Assignments
  for unit $i = 1, \ldots, I$ **do**
    for instance $n = 1, \ldots, N_i$ **do** ▷ N_i: pre-treatment unit-item instances for unit i
      Sample η_{in} using the collapsed Gibbs update in Eq. (S2.8)
      Update counts $n^u_{i\eta_{in}}$ and $n^f_{\eta_{in}, v_{in}}$
    end for
  end for
  (B) Posterior Draws for Factor Parameters
  for unit $i = 1, \ldots, I$ **do**
    Sample $\boldsymbol{\theta}_i$ as in Eq. (9) ▷ Unit–factor prevalences
  end for
  for factor $k = 1, \ldots, K$ **do**
    Sample $\boldsymbol{\phi}_k$ as in Eq. (9) ▷ Factor–item distributions
  end for
  (C) Item–Factor Associations and Construction of Outcomes
  for unit $i = 1, \ldots, I$ **do**
    for item $j = 1, \ldots, J$ **do**
      Sample z_{ij} using the posterior in Eq. (3) ▷ Item–factor assignments
    end for
    for factor $k = 1, \ldots, K$ **do**
      Compute factor-level outcomes Y_{ik} using Eq. (2)
      Compute pre-treatment features d_{ik} using Eq. (6)
    end for
  end for
  (D) Causal Regression Updates
  ▷ Aggregate ATE and HTE parameters
  Sample (α, τ, σ) and $(\boldsymbol{\alpha}^h, \boldsymbol{\tau}^h, \sigma^h)$ using Eqs. (S2.2)–(S2.5) with Y_i as the outcome
  ▷ Factor-level ATE and HTE parameters
  for factor $k = 1, \ldots, K$ **do**
    Sample $(\alpha_k, \tau_k, \sigma_k)$ using Eqs. (S2.2)–(S2.5) with Y_{ik}
    Sample $(\boldsymbol{\alpha}^h_k, \boldsymbol{\tau}^h_k, \sigma^h_k)$ using the heterogeneous specification
  end for
end for

S3 MCMC Convergence Diagnostics

We assess convergence of the MCMC algorithm used for latent-factor estimation and for the Bayesian decomposition of treatment effects. For the main specification with $K = 10$ latent

factors, Algorithm S2.1 was run for 2,000 iterations. Unless otherwise noted, the first 1,000 iterations are discarded as burn-in, and the remaining 1,000 posterior draws are retained for inference.

As an overall diagnostic, Figure S3.1 plots the trace of the average log-likelihood of the latent-factor model across sampling iterations. The chain stabilizes rapidly after the initial iterations and exhibits no discernible trends, drifts, or cyclical patterns. The trace displays behavior consistent with satisfactory mixing of the latent allocations and associated Dirichlet parameters.

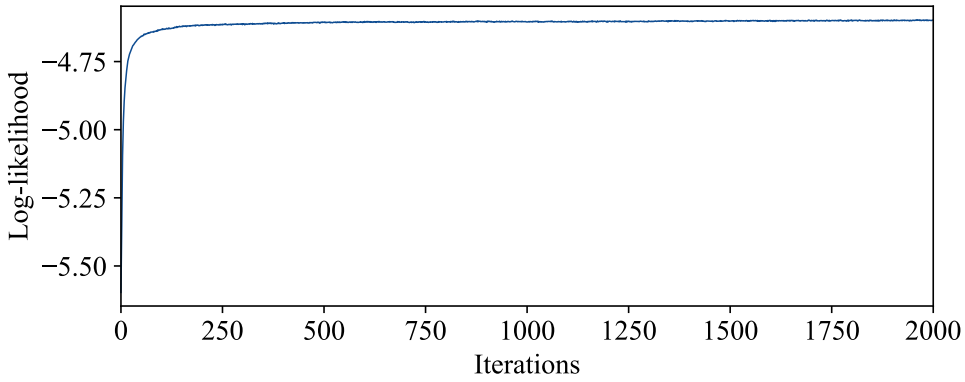


Figure S3.1: MCMC Draws for Average Log-Likelihood of the Latent-Factor Model

We next examine convergence for parameters central to the empirical analysis. Figure S3.2 reports the MCMC traces for the factor-level ATEs $\{\tau_k\}$ across the full 2,000 iterations. Each trajectory fluctuates around a stable center with no persistent drifts or signs of nonstationarity. The marginal variance appears roughly constant throughout the chain, and visual inspection reveals no stickiness or slow-moving segments indicative of poor mixing. Successive draws show no evidence evidence of severe autocorrelation that would impede effective sampling.

Based on these diagnostics, we discard the first half of the chain as burn-in and compute posterior summaries from the remaining draws, as reported in the main text. No additional convergence issues were detected for other parameter blocks (e.g., factor–item distributions, unit–factor prevalences, regression coefficients). All monitored chains exhibit behavior consistent with convergence to the stationary posterior distribution.

S4 Competing Factor Models

To assess whether pre-defined product categories provide an adequate representation of unit-item interactions, we compare our mixed-membership LDA specification to several fixed-membership and restricted benchmarks.

Let categories be indexed by $c = 1, \ldots, C$, and let item j belong to exactly one category, with within-category item distribution $\phi_c(j) = 1/J_c$ if $j \in c$ and 0 otherwise, where J_c denotes the number of items assigned to category c. Each unit i has category preferences $\boldsymbol{\theta}_i^c \sim \text{Dirichlet}(\boldsymbol{\alpha}^c)$, which govern the probability of selecting each product category. For

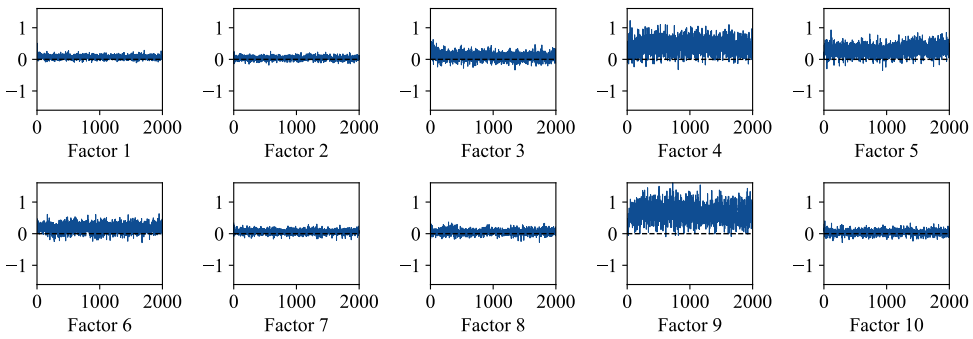


Figure S3.2: MCMC Draws for Factor-Level Treatment Effects

Model	Number of Factors	Perplexity (In)	Perplexity (Out)
Mixed-Membership LDA	10	98.53	144.70
Homogeneous LDA	10	282.67	292.68
Firm-Based Categories	5	2,602.90	2,641.12
Random Categories	5	2,998.53	3,086.16

Table S4.1: Perplexity Comparison Across Competing Factor Models

$n = 1, \ldots, N_i$, pre-treatment instances are generated as $c_{in} \sim \text{Categorical}(\boldsymbol{\theta}_i^c)$ and $v_{in} \sim \text{Categorical}(\phi_{c_{in}})$. Thus, each observation is generated by first sampling a category and then sampling an item uniformly within that category. This model enforces fixed item memberships across all units. Let $n_{ic} = \sum_{n=1}^{N_i} \mathbb{I}\{c_{in} = c\}$ denote the number of instances of unit i in category c. The posterior distribution of category preferences is $\boldsymbol{\theta}_i^c \mid \{c_{in}\} \sim \text{Dirichlet}(\alpha^c + n_{i1}, \ \ldots, \ \alpha^c + n_{iC})$.

We also estimate a restricted LDA variant in which factor memberships are homogeneous across all units. Specifically, we impose $\boldsymbol{\theta}_i = \boldsymbol{\theta}$ for all $i = 1, \ldots, I$, while maintaining the assumption that factors are probability distributions over items. Under this restriction, the posterior item–factor allocation satisfies $P(z_{ij} = k \mid \boldsymbol{\theta}, \boldsymbol{\phi}_k) \propto \theta_k \, \phi_{kj}$, and is identical across all units. We calibrate this model over a range of factor counts and evaluate fit using out-of-sample perplexity.

We compare predictive perplexity across mixed-membership LDA, homogeneous LDA, firm-based categories, and random category assignments of equal size. Results are reported in Table S4.1.

The firm-based category model substantially improves upon random category assignments, but remains orders of magnitude less predictive than latent factor models. Even the homogeneous LDA substantially outperforms firm categories, while the mixed-membership LDA achieves the best fit overall. These results indicate that pre-treatment consumer behavior is poorly represented by fixed product categories and is best described by a mixed-membership structure in which both underlying behavioral factors and item associations vary across units.

S5 Distribution of Latent Factor Associations Across Products

To assess the extent to which products exhibit multi-factor membership under the posterior distribution implied by Algorithm S2.1, we summarize the sampled item–factor assignments.

We first compute the assignment probabilities $P(z_{ij} = k \mid \boldsymbol{\theta}_i, \boldsymbol{\phi}_k)$ for all units i and items j in the pre-treatment data using the posterior mean estimates of $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_k$. Each unit-item pair (i, j) is then assigned to its most likely latent factor via $\hat{z}_{ij} = \arg\max_k \; P(z_{ij} = k \mid \boldsymbol{\theta}_i, \boldsymbol{\phi}_k)$. Finally, for each item j, we compute the number of unique factors to which it is assigned across all units, given by $U_j = \text{card}(\{\hat{z}_{ij} : i = 1, \ldots, I\})$.

Figure S5.1 reports the empirical distribution of $\{U_j\}$ across all products. Approximately 48% of products load predominantly on a single factor ($U_j = 1$), while the remaining 52% exhibit substantive posterior support on two or more factors. The substantial mass on $U_j \geq 2$ indicates that many products participate in multiple latent behavioral patterns, underscoring the value of a probabilistic latent-factor representation. Such cross-loading captures overlapping usage contexts and shared behavioral preferences that fixed administrative product categories are structurally unable to represent.

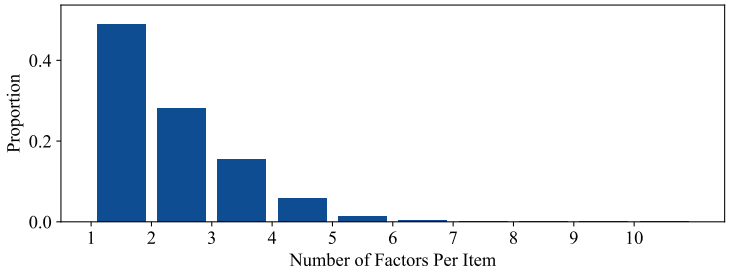


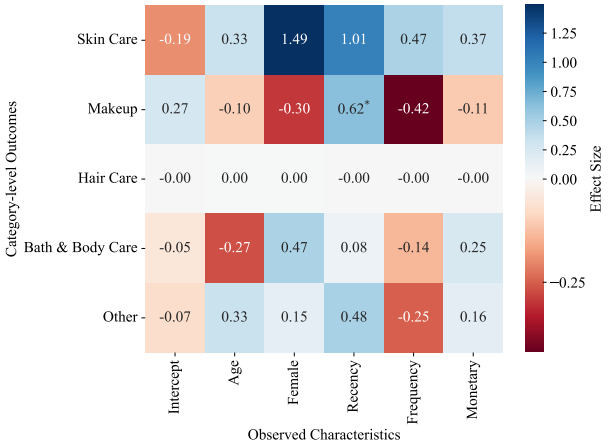
Figure S5.1: Distribution of Factor Count per Product

S6 Category-Level Heterogeneous Treatment Effects

For comparison with the latent-factor specification, we replicate the HTE analyses using the retailer's five administrative product categories as outcome partitions.

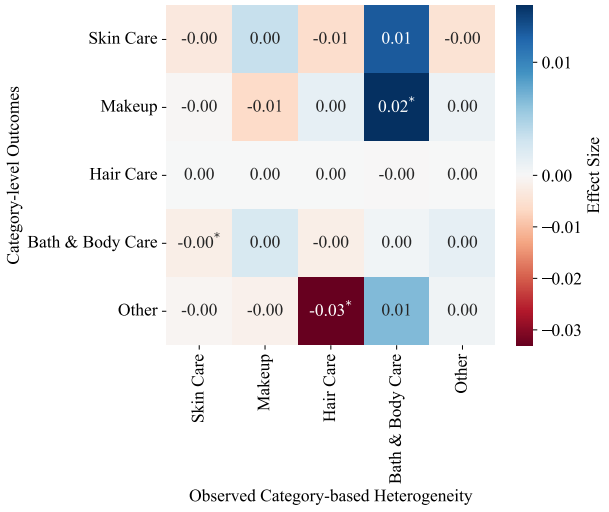
Figure S6.1 reports category-level HTEs by observed customer characteristics (gender, age, and RFM metrics). Across all categories, subgroup contrasts exhibit wide and overlapping 95% credible intervals. The few apparent differences are small in magnitude and not robust across categories, indicating substantially weaker discriminatory power relative to the latent-factor decomposition.

Figure S6.2 reports heterogeneity based on pre-treatment category-level spending. Reinforcement patterns are weak and inconsistent: diagonal coefficients are generally small and frequently include zero, and off-diagonal estimates do not exhibit any coherent substitution or complementarity structure. In contrast to the latent-factor results, the category-based specification fails to recover systematic patterns of heterogeneity.



Note: * indicates 95% Bayesian credible interval excludes 0.

Figure S6.1: Observed Heterogeneity in Category-Level Treatment Effects



Note: * indicates 95% Bayesian credible interval excludes 0.

Figure S6.2: Category Heterogeneity in Category-Level Treatment Effects

Overall, the category-based HTE analyses are statistically weaker and substantively less informative than their latent-factor counterparts. The absence of clear reinforcement or cross-category relationships supports the conclusion that fixed administrative taxonomies obscure much of the causal heterogeneity that the latent-factor representation is able to recover.

S7 Intervention Design Using Factor-Level Heterogeneity

This section illustrates how the proposed framework can be used to generate actionable diagnostics for intervention design and targeting.

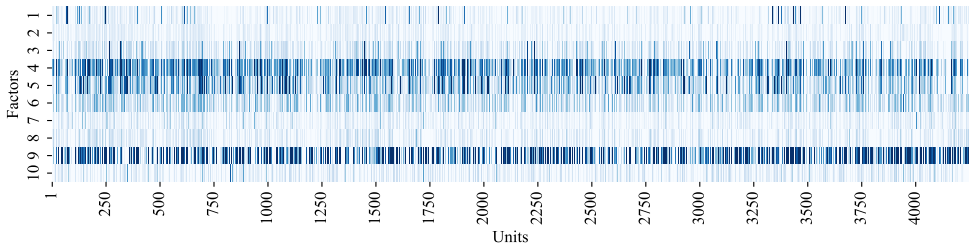
Beyond decomposing the aggregate treatment effect, our framework provides unit-factor-level HTE $\{\hat{\tau}_{ik}\}$. These capture how the intervention amplifies or attenuates outcomes along distinct latent behavioral dimensions for each unit, based on both observed covariates and latent pre-treatment factor exposures. As a result, the model identifies not only where treatment effects are strongest, but also where they are negative or misaligned with the intervention's objective.

Because unit-level treatment effects decompose additively across factors, the aggregate ATE satisfies

$$\tau \;=\; \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K}\tau_{ik} \;\;\leq\;\; \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K}\tau_{ik}\cdot\mathbb{I}\{\tau_{ik}\geq 0\},$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. This inequality highlights the potential gains from eliminating negative factor-level responses through targeted refinement.

Figure S7.1 displays the estimated unit-factor treatment effects $\hat{\tau}_{ik}$ in our empirical application. Approximately 25% of these effects are negative, indicating that a nontrivial share of unit-factor interactions respond adversely to the intervention. If these negative contributions were mitigated, the aggregate treatment effect would increase by approximately 19% relative to the observed benchmark.



Note: Darker colors indicate larger $\hat{\tau}_{ik}$.

Figure S7.1: Estimated Factor-level HTE

To illustrate at the unit level, consider a representative consumer with estimated total treatment effect $\hat{\tau}_i = 0.93$. This unit exhibits negative responses on four latent factors: Factor 8 (-0.332), Factor 5 (-0.140), Factor 9 (-0.033), and Factor 6 (-0.021). If these

negative contributions were neutralized, the unit's predicted response would increase to 1.46, implying a gain of 0.53 in incremental spending.

Crucially, the framework also identifies the specific items most strongly associated with each adverse unit-factor pair. For unit i and factor k, the items most responsible for the effect are those with the largest posterior association $p_{ijk} \propto \theta_{ik}\,\phi_{kj}$. In this example, the most influential items are {1520, 318} in Factor 5, {114, 847} in Factor 6, {595, 128} in Factor 8, and {1352, 1516} in Factor 9. These diagnostics illustrate how the latent-factor decomposition translates directly into intervention design and targeting strategies by linking heterogeneous treatment responses to specific behavioral mechanisms and actionable items.