Estimating Heterogeneous Treatment Effects in Online Advertising Experiments: A Hierarchical Bayesian Approach

Aljoša Vodopija
Teads
Ljubljana, Slovenia
aljosa.vodopija@teads.com

Anže Alič
Teads
Ljubljana, Slovenia
anze.alic@teads.com

Tim Poštuvan
Teads
Ljubljana, Slovenia
tim.postuvan@teads.com

Martin Jakomin
Teads
Ljubljana, Slovenia
martin.jakomin@teads.com

Blaž Škrlj
Teads
Ljubljana, Slovenia
blaz.skrlj@teads.com

Abstract

We present a production hierarchical Bayesian framework for estimating heterogeneous treatment effects in large-scale online advertising experiments. By jointly modeling campaign-level effects and cross-campaign heterogeneity, it addresses a limitation of inverse-variance weighting, the standard meta-analytic baseline, which becomes unstable in such challenging regimes. Across simulations spanning a range of effect sizes and heterogeneity levels, the approach delivers more reliable inference under high heterogeneity while matching or improving estimation accuracy overall. We further demonstrate its practical value in a production demand-side platform handling over 5 million requests per second, integrating it with sequential testing and evaluating it across three experiments covering over 35,000 campaigns from 3,000 advertisers.

CCS Concepts

• **Information systems** → **Online advertising**; **Recommender systems**; • **Mathematics of computing** → *Bayesian computation.*

Keywords

Hierarchical Bayesian models, Meta-analysis, A/B testing, Sequential testing, Online advertising, Treatment effect heterogeneity

ACM Reference Format:
Aljoša Vodopija, Anže Alič, Tim Poštuvan, Martin Jakomin, and Blaž Škrlj. 2026. Estimating Heterogeneous Treatment Effects in Online Advertising Experiments: A Hierarchical Bayesian Approach. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26).* ACM, New York, NY, USA, 6 pages.

1 Introduction

Performance advertising has shifted from click-based to conversion-based objectives, making conversion rate (CVR) and cost per acquisition (CPA) central metrics in modern advertising systems [9]. Yet interpreting experiments built around these metrics is notoriously difficult. Conversions are sparse and diverse, ranging from low-intent page visits to high-value purchases. This diversity inflates the variance of observed CPAs and introduces genuine heterogeneity in treatment effects across campaigns, driven by differences in targeting, creatives, and auction dynamics [4].

Despite its practical importance, the academic literature on CPA experimentation remains thin. Existing approaches typically treat all conversions as equivalent or report results at coarse aggregation levels such as advertiser vertical. A notable exception is the meta-analytical framework of [12], which estimates the overall treatment effect on return on investment (ROI) by treating each campaign as an independent experimental unit—a natural choice, since goals and targeting are defined at the campaign level. Campaign-level effects are then aggregated using inverse-variance weighting (IVW) to account for uncertainty and cross-campaign diversity.

We build on this line of work but focus on the count-based nature of CPA rather than ROI. A well-known limitation of IVW in this setting is its instability under high effect heterogeneity and sparse conversion counts—precisely the regime in which conversion experiments operate [2, 13]. That is why we turn to Hierarchical Bayesian models, which naturally accommodate heterogeneity and remain stable under sparse counts. Although such models and Bayesian testing are well-established in statistics [3], to our knowledge, they have not been combined for aggregating treatment effects in online advertising experiments. We close this gap with a framework designed for production-scale deployment:

- We adapt Bayesian modeling to CPA experimentation, jointly modeling campaign-level effects and cross-campaign variability through a hierarchical model that remains robust under sparse conversions and strong effect heterogeneity.
- We operationalize the framework within a Bayesian sequential testing pipeline based on highest density intervals (HDIs) and regions of practical equivalence (ROPE), enabling practical early-stopping decisions in production.
- The framework is running in production on a large-scale demand-side platform (DSP) spanning over 35,000 campaigns and more than 3,000 advertisers, driving high-stakes decisions on platform-wide model rollouts. We validate it through calibrated simulations and demonstrate it on three case studies: an A/A validation, a CVR model rollout, and a creative retrieval model rollout.

KDD '26, Jeju, Korea

2 Background and framework formulation

We introduce a statistical framework for analyzing online experiments conducted across multiple advertising campaigns. The framework formalizes how treatment effects are estimated at the campaign level and subsequently aggregated using meta-analytic principles, explicitly accounting for uncertainty and cross-campaign heterogeneity.

2.1 Data and basic definitions

A typical dataset in our setup, $\mathcal{D}$, obtained from A/B tests consists of m advertisers running a total of n campaigns. Each campaign i is evaluated under two variants: control ($j = 0$) and treatment ($j = 1$). For each variant, we observe campaign spend, s_{ij}, and the number of conversions, c_{ij}.

A commonly used performance metric, indicating the campaign's success, is CPA, expressed as $\mathrm{CPA}_{ij} = s_{ij}/c_{ij}$. For statistical modeling, it is more convenient to work with its inverse, the conversion intensity (CVI). Under a Poisson model with spend as an offset, its maximum-likelihood estimate is $\mathrm{CPA}_{ij}^{-1} = c_{ij}/s_{ij}$.

2.2 Inverse-variance weighting

Following [12], we treat each campaign as an independent experimental unit. For each campaign i, the treatment effect is the log rate ratio between treatment and control, with a continuity correction [5] of 0.5 added to each count to handle zeros:

$$\hat{\theta}_i = \log\left(\frac{(c_{i1}+0.5)/s_{i1}}{(c_{i0}+0.5)/s_{i0}}\right), \quad \mathrm{SE}_i = \sqrt{\frac{1}{c_{i0}+0.5} + \frac{1}{c_{i1}+0.5}}. \tag{1}$$

The log transformation ensures symmetry between multiplicative gains and losses, and stabilizes variance. The standard error follows from the Poisson assumption with spend treated as a fixed offset.

As a baseline, we use IVW, a standard meta-analytic approach. The global treatment effect is estimated as

$$\hat{\theta} = \frac{\sum_{i=1}^{n} w_i \hat{\theta}_i}{\sum_{i=1}^{n} w_i}, \quad w_i = \frac{1}{\mathrm{SE}_i^2}, \tag{2}$$

assigning greater influence to campaigns with lower estimation uncertainty, which here corresponds to campaigns with higher conversion counts.

To account for treatment effect heterogeneity, we adopt a random-effects IVW formulation in which the campaign-level true effects are drawn from a common distribution,

$$\theta_i \sim \mathrm{Normal}(\theta, \tau^2), \quad \hat{\theta}_i \mid \theta_i \sim \mathrm{Normal}(\theta_i, \mathrm{SE}_i^2), \tag{3}$$

implying the marginal $\hat{\theta}_i \sim \mathrm{Normal}(\theta, \mathrm{SE}_i^2 + \tau^2)$ and updated weights

$$w_i = \frac{1}{\mathrm{SE}_i^2 + \tau^2}, \tag{4}$$

where τ is estimated using the Paule–Mandel method [10], a standard method-of-moments estimator for the between-study variance in random-effects meta-analysis.

3 Methods

We propose a Bayesian alternative to IVW. Rather than aggregating per-campaign point estimates, we fit a single hierarchical model jointly across all campaigns, and use the resulting posterior over the global treatment effect for sequential testing.

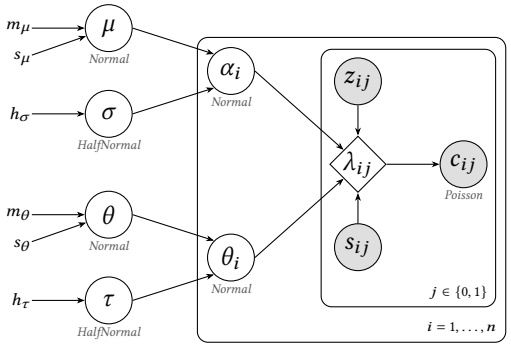

Figure 1: Plate diagram of the HBM. Unshaded circles are latent variables, shaded circles are observed, the diamond is a deterministic node, and bare symbols are fixed constants.

3.1 Hierarchical Bayesian model

We model conversions across all campaigns jointly using a hierarchical Bayesian model (HBM) based on Poisson regression, summarized as a plate diagram in Figure 1. The full specification proceeds in three layers: a likelihood linking observed conversions to latent rates, priors over campaign-level parameters, and hyperpriors over the population-level parameters.

Likelihood. For each campaign i and variant $j \in \{0, 1\}$, the observed conversion count c_{ij} follows a Poisson distribution with mean λ_{ij}, as in [1]:

$$c_{ij} \sim \mathrm{Poisson}(\lambda_{ij}), \quad \log(\lambda_{ij}) = \alpha_i + \theta_i z_{ij} + \log(s_{ij}), \tag{5}$$

where $z_{ij} \in \{0, 1\}$ is the treatment indicator and $\log(s_{ij})$ is a Poisson offset representing exposure. The campaign-level parameters have a direct interpretation: α_i is the log CVI on the control variant, and θ_i is the log rate ratio between treatment and control.

The model assumes conditionally Poisson conversion counts. In practice, advertising conversions may exhibit overdispersion or zero inflation due to unmodeled campaign dynamics. However, modeling spend as an offset mitigates part of this variability by normalizing conversion counts relative to campaign scale. The Poisson likelihood is further justified empirically in Section 4.2.3.

Priors. Campaign-level parameters are drawn from population-level distributions,

$$\alpha_i \sim \mathrm{Normal}(\mu, \sigma^2), \quad \theta_i \sim \mathrm{Normal}(\theta, \tau^2). \tag{6}$$

Here μ is the average log CVI on the control variant across campaigns, and θ is the average log rate ratio between treatment and control—the global treatment effect. The scale parameters σ and τ govern between-campaign heterogeneity, with σ controlling the spread of baseline CVIs and τ controlling the spread of treatment effects. This induces *partial pooling*: campaigns with many conversions contribute precise information to the posteriors of the population parameters, while low-volume campaigns are shrunk toward the population mean by an amount that depends on their individual uncertainty.

Hyperpriors. The population-level parameters receive weakly informative hyperpriors following [11]:

$$\begin{aligned} \mu &\sim \text{Normal}(m_\mu, s_\mu^2), \quad \sigma \sim \text{HalfNormal}(h_\sigma), \\ \theta &\sim \text{Normal}(m_\theta, s_\theta^2), \quad \tau \sim \text{HalfNormal}(h_\tau), \end{aligned} \tag{7}$$

where $m_\mu, s_\mu, h_\sigma, m_\theta, s_\theta, h_\tau$ are fixed constants. This additional layer lets the data determine both the global effect and the degree of pooling, while the weakly informative form ensures the priors do not dominate the posterior. The selection of the constants is described in Section 4.1.5.

The hierarchical model differs qualitatively from IVW: while IVW produces a weighted average of fixed per-campaign point estimates $\hat{\theta}_i$, HBM jointly estimates the per-campaign effects and the population distribution they are drawn from, with information flowing in both directions. The global treatment effect θ and the between-campaign heterogeneity τ are recovered directly from the posterior. In particular, τ is a parameter of the model rather than a fixed point estimate, so uncertainty in heterogeneity is propagated into the posterior of θ.

3.2 Sequential testing

Sequential testing integrates naturally with Bayesian posterior updating, allowing the treatment-effect posterior to be recomputed incrementally as new data arrive. We adopt the decision procedure from [8], based on the HDI of the posterior and a ROPE. The ROPE, denoted $[-\delta, \delta]$, is a range of values of θ considered operationally indistinguishable from the null hypothesis.

The choice of δ is domain-specific and discussed in Section 4.1.4. Anchored to the aggregate A/A noise floor, δ is stable and largely insensitive to campaign mix in large-scale experiments spanning many advertisers, but for narrow breakdowns (e.g., a single advertiser vertical) the noise floor differs, so δ must be recalibrated per breakdown rather than borrowed across experiment types.

Unlike alternatives such as mSPRT [7] or Bayes factors [3], this procedure expresses decisions in units of θ and separates statistical detection from practical significance. It lacks formal optional-stopping guarantees, but Section 4.1.4 shows empirically that the false-decision rate stays controlled under hourly monitoring over a one-week horizon.

At each monitoring time (e.g., hourly), we compare the $1-\alpha$ HDI of the posterior of θ against the ROPE and take one of three actions, the first two of which are early-stopping criteria:

- **Declare an effect** if the HDI lies entirely outside the ROPE.
- **Declare practical equivalence** if the HDI lies entirely within the ROPE.
- **Continue** if the HDI overlaps the ROPE boundary.

The monitoring procedure terminates as soon as the first or second condition is met. To prevent indefinite running when the posterior concentrates inside neither region, we additionally apply a precision-stopping criterion: we stop when the HDI width falls below a threshold $w_{\max}$, whose selection is discussed in Section 5.1. At that point, the interval is precise but overlaps the ROPE boundary, so neither conclusion is justified, and further data are unlikely to resolve the ambiguity.

Table 1: Data-generating process parameters and their values used in offline experiments. Distribution families are indicated in parentheses when not specified in the text.

Symbol	Description	Value
m	Number of advertisers	3,000
λ_c	Campaigns per advertiser (1 + Poisson)	9
μ_c, σ_c	CPA mean, std. (LogNormal)	50, 20
σ_w	Within-advertiser CPA spread (Normal)	5
μ_s, σ_s	Spend mean, std. (LogNormal)	160, 200
ϕ	Dispersion parameter	120
$\theta^\star$	Treatment effect mean	0 / 1.5δ / 2δ / 3δ
$\tau^\star$	Treatment effect heterogeneity	1δ / ... / 10δ

4 Offline experiments

Our offline evaluation considers two representative scenarios designed to reflect common challenges in practice: small effect sizes and high effect heterogeneity.

4.1 Setup

4.1.1 Data-generating process and calibration. To benchmark methods against ground truth, we simulate datasets that mirror the structure of production A/B tests and produce data with a known global treatment effect $\theta^\star$ and treatment effect heterogeneity $\tau^\star$. The data-generating process (DGP) draws advertisers, campaigns within advertisers, and conversion counts within campaigns.

For each advertiser, we sample the number of campaigns and an advertiser-level CPA. Then, within each campaign, we sample campaign-level CPA_i, spend s_i (equal for control and treatment variants), and treatment effect $\theta_i^\star \sim \text{Normal}(\theta^\star, \tau^{\star 2})$. The specific distributions and values are listed in Table 1. Conversion counts are then drawn as

$$c_{ij} \sim \text{NegBin2}(\lambda_{ij}, \phi), \quad \log(\lambda_{ij}) = \alpha_i^\star + \theta_i^\star z_{ij} + \log(s_i), \tag{8}$$

where $\alpha_i^\star = -\log(\text{CPA}_i)$ is the log baseline CVI and $z_{ij} \in \{0, 1\}$ is the treatment indicator. We sample from NegBin2 to allow simulated counts to exhibit overdispersion relative to the Poisson inference model, with $\phi > 0$ controlling overdispersion through the variance relation $\text{Var}(c_{ij}) = \lambda_{ij} + \lambda_{ij}^2/\phi$.

The advertiser–campaign hierarchy in the DGP also induces within-advertiser dependence between campaigns, which the inference model does not explicitly account for. Together, the NegBin2 sampling and the advertiser hierarchy ensure that the DGP departs from the inference model along practically relevant dimensions, allowing us to assess robustness under realistic misspecification. We calibrated the DGP to five days of production data; the resulting parameters, slightly shifted to avoid disclosing production figures, appear in Table 1.

4.1.2 Evaluation protocol. We compare HBM with IVW on type I error, power, interval coverage, and point-estimate accuracy (bias and RMSE of $\hat{\theta}$), all at nominal $\alpha = 5\%$.

To evaluate type I error and power consistently, each method needs a decision rule for declaring a positive effect. For IVW we use a one-sided test based on the asymptotic normality of the estimator, rejecting when the p-value falls below α. For HBM we declare a

positive effect when $P(\theta > 0 \mid \mathcal{D}) > 1 - \alpha$, the Bayesian one-sided analogue at the same nominal level.

We benchmark against IVW because it is a standard production baseline for aggregating campaign-level effects in online advertising experimentation [12], making it the most natural point of comparison for a hierarchical alternative like ours. More sophisticated alternatives, including generalized linear mixed models (GLMMs), are left for future work.

4.1.3 Implementation. All methods are implemented in Python. We provide a custom package[1] that covers the full experimental pipeline—data simulation, model fitting, and evaluation—to ensure reproducibility and to enable application to proprietary datasets.

For IVW, the between-campaign heterogeneity τ is estimated via the Paule–Mandel method [10], using the meta-analysis utilities in `statsmodels`. HBM is implemented in `PyMC` and fit using its built-in No-U-Turn Sampler (NUTS) [6] with 4 chains, 5,000 warm-up iterations, and 3,000 posterior samples per chain at a target acceptance rate of 0.95. We chose these settings based on preliminary experiments to ensure reliable convergence. All reported runs passed standard convergence diagnostics for μ, σ, θ, and τ.

Inference for a typical hourly snapshot involving approximately 30,000 campaigns completed within 7–9 minutes on 4 CPU cores. Scaling is not strictly linear in the number of campaigns, but it remains feasible; we observed 45–52 minutes at 100,000 campaigns on the same hardware. At larger scales, variational inference or stochastic-gradient HMC offer approximate alternatives.

4.1.4 Reference scale calibration. We report effects in units of a reference scale δ, calibrated *a priori* from a set of historical week-long A/A tests run on our production DSP prior to the experiments analyzed in this paper. Specifically, δ is set to 1.2× the maximum absolute 95% HDI endpoint observed across these A/A tests. We set the 1.2 multiplier in consultation with product stakeholders, balancing conservativeness with the smallest practically meaningful effect. Anchoring results to δ makes them directly interpretable for experiment-design decisions.

We verified this choice empirically. Across 200 simulated week-long runs with hourly monitoring, the rate of false effect declarations was 0.0% under the strict null ($\theta^\star = 0$) and 3.7% under the practical null ($\theta^\star = 0.9\delta$), both below the nominal 5% level.

Expressing effects in units of δ also makes the methodology portable: δ can be recalibrated from any platform's A/A history, leaving hyperprior scales and decision thresholds unchanged.

4.1.5 Hyperpriors. The HBM hyperpriors are chosen to be weakly informative on the δ-based scale, allowing the data to dominate while ruling out implausibly large effects. For the control-variant log CVI, we set m_μ to the historical mean log CVI observed on our platform and use $s_\mu = 10\delta$. For the global treatment effect, we use $m_\theta = 0$ and $s_\theta = 5\delta$, centering the prior at the null while remaining wide enough to accommodate any practically relevant effect. The heterogeneity scales receive half-normal priors with $h_\sigma, h_\tau = 10\delta$.

Sensitivity checks varying $s_\mu, s_\theta \in \{5\delta, 7.5\delta, 10\delta\}$ and $h_\sigma, h_\tau \in \{5\delta, 10\delta, 15\delta\}$ shifted posterior means of θ by less than 0.02δ and HDI widths by less than 2% across simulation settings, confirming that inferences are insensitive to reasonable hyperprior variations.

[1] https://github.com/outbrain-inc/ab-tea-lab

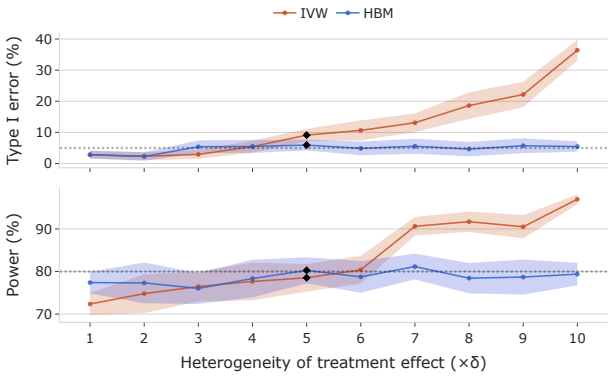

Figure 2: Type I error (top, $\theta^\star = 0$) and power (bottom, $\theta^\star = 2\delta$) as a function of treatment effect heterogeneity $\tau^\star$, over 200 replications. Shaded bands show ±1 standard error. Black diamonds mark the $\tau^\star = 5\delta$ slice reported in Table 2. Dashed lines indicate the nominal $\alpha = 5\%$ and the 80% power target.

4.2 Results

4.2.1 Effect sizes. Table 2 reports empirical measures across various effect sizes at fixed heterogeneity $\tau^\star = 5\delta$, a level representative of what we typically observe in our production data. Results are averaged over 200 simulation runs. Under the null, HBM controls type I error near the nominal level (5.9% vs. $\alpha = 5\%$), while IVW is anti-conservative at 9.1%. This reflects the difficulty of estimating $\tau^\star$ from a single experiment with moment-based estimators such as Paule–Mandel, where underestimating $\tau^\star$ shrinks standard errors and inflates rejections [13]. HBM avoids this by jointly inferring θ and τ and propagating uncertainty through the posterior.

Power differences are within ±2.2% across non-null settings, with HBM being slightly ahead at $\theta^\star = 2\delta$ and IVW at $\theta^\star = 3\delta$. These IVW figures should be read alongside its inflated type I error: part of its apparent sensitivity comes from a more permissive rejection rule rather than genuine efficiency. Coverage is similar: both methods are close to the nominal 95%, with HBM closer under the null.

HBM achieves lower RMSE than IVW at all effect sizes except $\theta^\star = 3\delta$, with the largest gains at small and moderate effects—the regime in which A/B tests typically operate. Bias differences are more pronounced: IVW shows a positive bias of 0.32δ under the null and 0.18δ at $\theta^\star = 1.5\delta$, while HBM remains nearly unbiased ($\leq 0.09\delta$) throughout.

4.2.2 Effect heterogeneity. Figure 2 shows the most consequential difference between the methods. Under the null, HBM holds type I error at the nominal 5% across the full range of $\tau^\star$, while IVW is well-calibrated only for $\tau^\star \leq 4\delta$ and then deteriorates rapidly, exceeding 35% at $\tau^\star = 10\delta$. While $\tau^\star \approx 5\delta$ is typical across our broader production data, smaller and more diverse advertiser subpopulations routinely exhibit substantially higher heterogeneity, well into the range where IVW becomes unreliable.

The power panel mirrors this pattern. Both methods are close to the 80% design target at $\tau^\star = 5\delta$ (HBM 80.3%, IVW 78.5%), but as $\tau^\star$ grows HBM remains stable while IVW power climbs above 95%—an artifact of its inflated rejection rate. At low heterogeneity ($\tau^\star \leq 2\delta$),

Table 2: Simulation results comparing IVW and HBM across different true treatment effects $\theta^\star$, at fixed heterogeneity $\tau^\star = 5\delta$. Testing performance is reported in percent, and point estimation in units of δ.

	$\theta^\star = 0$		$\theta^\star = 1.5\delta$		$\theta^\star = 2\delta$		$\theta^\star = 3\delta$	
Measure	IVW	HBM	IVW	HBM	IVW	HBM	IVW	HBM
Type I error	9.10	**5.90**	–	–	–	–	–	–
Power	–	–	**67.00**	66.30	78.50	**80.30**	**97.70**	95.50
Coverage	93.50	**95.20**	**97.00**	96.00	**98.20**	96.30	**95.70**	94.20
RMSE	0.92	**0.84**	0.94	**0.82**	0.81	**0.75**	**0.79**	0.84
Bias	0.32	**0.02**	0.18	**0.02**	−0.04	**−0.02**	−0.19	**−0.09**

IVW is mildly under-powered, suggesting it is also conservative when $\tau^\star$ is small. In short, HBM provides reliable inference across the full range of heterogeneity, while IVW's validity is restricted to low values of $\tau^\star$.

4.2.3 Robustness to overdispersion. The Poisson likelihood in Eq. 5 is less sensitive to overdispersion than one might expect, since the spend offset regulates variance. Although we observe mild overdispersion in practice, switching to NegBin2 on our calibrated simulations shifted posterior means of θ by less than 0.02δ and HDI widths by less than 2.5%, while raising inference time from ∼8 to ∼12 minutes per snapshot. We therefore retain Poisson, noting that Eq. 5 can be swapped for NegBin2 with a prior on ϕ if overdispersion becomes material elsewhere.

5 Online experiments

In this section, we present three online case studies demonstrating the application of the proposed method in a production setting. Results are summarized in Figure 3.

5.1 Common setup

Studies I and III are run on the full eligible inventory while Study II is restricted to a smaller advertiser population (see Section 5.3). All three experiments are conducted on our production DSP and share the following protocol. Randomization is performed at the auction (request) level via a spend-based allocation mechanism, where each campaign distributes half of its budget to the control variant and half to the treatment variant. We exclude internal testing campaigns, whose objectives do not reflect real advertiser behavior, and campaigns with cumulative spend below $1, as they provide unreliable signals and violate the equal-spend assumption.

Each experiment runs for approximately five consecutive days, with posterior means and 95% HDIs computed hourly and evaluated against a fixed ROPE. Sequential testing begins six hours after experiment launch, since earlier data is insufficient for reliable posterior convergence under the hierarchical model. The ROPE half-width δ is calibrated as described in Section 4.1.4, and the precision-stopping threshold is set to $w_{\max} = \delta$. This calibration yielded stable operational behavior across repeated experiments.

5.2 Case study I: Validation via an A/A test

The hypothesis under test is equivalence between control and treatment. The decision rule requires the HDI of the posterior effect to fall entirely within the ROPE, validating both the correctness of the hierarchical model and the proposed sequential testing procedure.

Figure 3a shows the sequential posterior of the treatment effect for CVI. The posterior mean converges toward zero and the 95% HDI falls entirely within the ROPE, correctly leading to an *equivalence* decision. The HDI enters the ROPE at approximately 72 hours, with the precision criterion satisfied around 92 hours. An analogous experiment on click intensity (defined as the inverse of CPC) reached the equivalence decision within roughly 24 hours.

5.3 Case study II: New CVR prediction model

Approximately 500 advertisers running about 7,000 campaigns participate in the experiment. This relatively small population also lets us assess the robustness of the proposed approach under a limited number of experimental units. The treatment variant uses a new CVR model developed offline, with additional machine learning features and hyperparameter optimization, while the control variant uses the production model. We test for a performance lift, i.e., an increase in CVI. The decision rule requires the HDI of the posterior effect to lie entirely above the ROPE.

Figure 3b shows a clear and sustained lift in CVI for the treatment variant. The posterior mean stabilizes around $+2\delta$, and the 95% HDI clears the upper edge of the ROPE within approximately 32 hours, triggering an early-stopping decision in favor of the treatment well before the planned five-day horizon. The lift remains stable for the remainder of the experiment, supporting the deployment of the new CVR model. Notably, despite the relatively small advertiser population, the large magnitude of the lift was sufficient for the framework to reach a confident decision, consistent with the offline results in Section 4.2.1.

5.4 Case study III: New creative retrieval model

The treatment variant deploys a new two-tower retrieval model used in creative pre-ranking, obtained through hyperparameter optimization of the production model. As in the previous case study, we test for a practical lift in CVI, requiring the HDI of the posterior effect to lie entirely above the ROPE.

Figure 3c shows a small positive effect for the treatment, with the posterior mean stabilizing around $+\delta$. The 95% HDI overlaps the upper boundary of the ROPE throughout the experiment, so the early-stopping criterion is never triggered. The precision criterion is met at approximately 72 hours, providing sufficient evidence that the effect is too small to be practically significant. The decision is

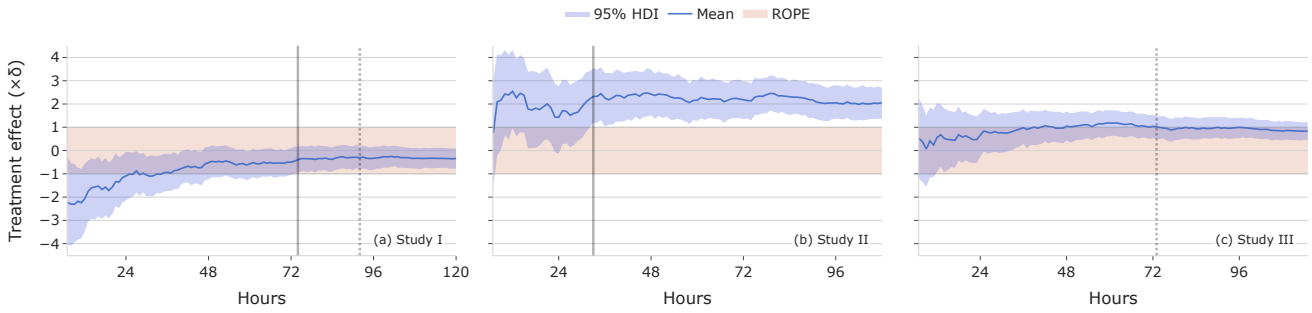

Figure 3: Sequential Bayesian testing for three case studies: (a) A/A test, (b) CVR model evaluation, and (c) retrieval model evaluation. Posterior means (solid blue line) and 95% HDIs (blue band) are updated hourly against a fixed ROPE (red band). The solid vertical line marks the earliest early-stopping point, and the dotted line indicates when the HDI width drops below $w_{\max}$.

robust to δ: shrinking the ROPE half-width to 0.7δ never lets the HDI clear it, so the no-deployment recommendation holds across a range far wider than the ±5% A/A-to-A/A drift in δ calibration.

Without the framework, the small positive posterior mean ($+\delta$) would have been interpreted as a directional win and released to production. Both conventional decision paths reach statistical significance well within the experiment: the 95% HDI lies entirely above zero after approximately 30 hours (Figure 3c), and the IVW one-sided test rejects the null after approximately 25 hours.

The HDI/ROPE rule separates statistical from practical significance, recommending against deployment because the effect, even if real, is too small to justify the operational costs. It thereby prevented a low-value deployment that conventional significance tests would have passed—the framework's most consequential property.

6 Conclusions

We proposed a hierarchical Bayesian framework for estimating heterogeneous treatment effects in online advertising experiments, combined with a sequential testing procedure based on HDIs and a ROPE that enables earlier decisions expressed in business-relevant terms. We benchmarked it against IVW on calibrated simulations and validated it in production.

Our results show that HBM is most beneficial when treatment-effect heterogeneity is high. When heterogeneity is moderate, IVW remains a competitive and computationally lighter alternative, so the choice between the two methods should be guided by the expected effect heterogeneity regime. The benefit of the hierarchical approach is therefore not raw sensitivity, but more robust and interpretable inference under the sparse and heterogeneous conversion regimes commonly encountered in online advertising.

This work has two main limitations that motivate future research. First, campaigns from the same advertiser are not independent. Our model omits this dependence, and our simulations inherit it: campaign effects are drawn from a single common Normal matching the inference prior in Eq. 6, so the data-generating process and inference model agree by construction and may understate real heterogeneity. Advertiser-level variation and heavy-tailed effects thus remain untested; an additional hierarchical level above the campaign priors could address both. Second, the comparison should extend beyond IVW to GLMMs, and time-to-decision could be shortened with informative priors built from historical control data.

Acknowledgments

We would like to thank Alberto Donizetti, Andraž Tori, and Robert Dupuy for valuable brainstorming discussions and insightful comments. During manuscript preparation, Claude (Opus 4.7, Anthropic) was used to assist with language editing. We reviewed the outputs and take full responsibility for the final content.

References

[1] Pantelis G Bagos and Georgios K Nikolopoulos. 2009. Mixed-effects Poisson regression models for meta-analysis of follow-up studies with constant or varying durations. *The International Journal of Biostatistics* 5, 1, Article 21 (2009). doi:10.2202/1557-4679.1168

[2] Winston Chou, Cameron Gray, Nathan Kallus, Aurélien Bibaut, and Simon Ejdemyr. 2025. Evaluating Decision Rules Across Many Weak Experiments. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*. ACM. doi:10.1145/3711896.3737217

[3] Alex Deng, Jiannan Lu, and Shouyuan Chen. 2016. Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing. In *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, Montreal, QC, Canada, 243–252. doi:10.1109/DSAA.2016.33

[4] Peter Ebbes, Eva Ascarza, and Oded Netzer. 2026. *Learning from Many Experiments: A Hierarchical Bayesian Framework for Decomposing Marketing Treatment Heterogeneity.* Working Paper hal-05562965. HEC Paris, Harvard Business School, and Columbia Business School. repec.org

[5] John J. Gart and James R. Zweifel. 1967. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* 54, 1-2 (1967), 181–187.

[6] Matthew D. Hoffman and Andrew Gelman. 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15, 47 (2014), 1593–1623.

[7] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2022. Always Valid Inference: Continuous Monitoring of A/B Tests. *Operations Research* 70, 3 (2022), 1806–1821. doi:10.1287/opre.2021.2135

[8] John K. Kruschke. 2015. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2 ed.). Academic Press, Boston.

[9] Weitong Ou, Bo Chen, Xinyi Dai, Weinan Zhang, Weiwen Liu, Ruiming Tang, and Yong Yu. 2024. A Survey on Bid Optimization in Real-Time Bidding Display Advertising. *ACM Transactions on Knowledge Discovery from Data* 18, 3, Article 62 (2024), 31 pages. doi:10.1145/3628603

[10] Robert C. Paule and John Mandel. 1982. Consensus values and weighting factors. *J. Res. Nat. Bur. Standards* 87, 5 (1982), 377–385.

[11] Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. 2014. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 70, 4 (2014), 1023–1032.

[12] Shahriar Shariat, Burkay Orten, and Ali Dasdan. 2015. Online Model Evaluation in a Large-Scale Computational Advertising Platform. In *2015 IEEE International Conference on Data Mining (ICDM)*. IEEE, 369–378. doi:10.1109/ICDM.2015.32

[13] Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian P T Higgins, Dean Langan, and Georgia Salanti. 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods* 7, 1 (2016), 55–79. doi:10.1002/jrsm.1164