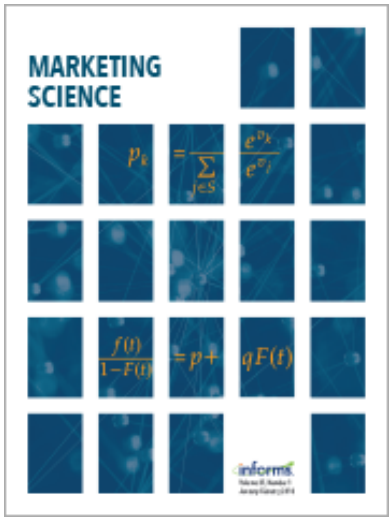
This article was downloaded by: [216.73.217.10] On: 05 November 2026, At: 15:00
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA



Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Overall Evaluation Criteria for A/B Tests: A Meta-Analytic Approach

Easton Huch, Stephanie A. Nail, Terran Melconian, Xan Vongsathorn, Fred Feinberg

To cite this article:
Easton Huch, Stephanie A. Nail, Terran Melconian, Xan Vongsathorn, Fred Feinberg (2026) Overall Evaluation Criteria for A/B Tests: A Meta-Analytic Approach. Marketing Science

Published online in Articles in Advance 05 Nov 2026

. https://doi.org/10.1287/mksc.2025.0164

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. You are free to download this work and share with others, but cannot change in any way or use commercially without permission, and you must attribute this work as "*Marketing Science*. Copyright © 2026 The Author(s). https://doi.org/10.1287/mksc.2025.0164, used under a Creative Commons Attribution License: https://creativecommons.org/licenses/by-nc-nd/4.0/."

Copyright © 2026 The Author(s)

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–23
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Overall Evaluation Criteria for A/B Tests: A Meta-Analytic Approach

Easton Huch,[a,*] Stephanie A. Nail,[b] Terran Melconian,[c] Xan Vongsathorn,[c] Fred Feinberg[d]

[a] Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109; [b] Applied Science and Data Science, Zillow Experimentation Platform, Zillow, Seattle, Washington 98101; [c] Applied Science, Zillow Experimentation Platform, Zillow, Seattle, Washington 98101; [d] Stephen M. Ross School of Business and Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109

*Corresponding author

ehuch@jhu.edu (E. Huch), stephn@zillowgroup.com (S. A. Nail), terranm@zillowgroup.com (T. Melconian), xanv@zillowgroup.com (X. Vongsathorn), feinf@umich.edu (F. Feinberg)

https://orcid.org/0000-0002-3822-1883 (E. Huch), https://orcid.org/0000-0003-2238-0721 (F. Feinberg)

Received: April 7, 2025
Revised: December 9, 2025; April 20, 2026; June 9, 2026
Accepted: June 12, 2026
Published Online in Articles in Advance: November 5, 2026

https://doi.org/10.1287/mksc.2025.0164

Copyright: © 2026 The Author(s)

Open Access Statement: This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. You are free to download this work and share with others, but cannot change in any way or use commercially without permission, and you must attribute this work as "*Marketing Science*. Copyright © 2026 The Author(s). https://doi.org/10.1287/mksc.2025.0164, used under a Creative Commons Attribution License: https://creativecommons.org/licenses/by-nc-nd/4.0/."

Abstract. A/B tests in digital marketing and application development typically use short-term metrics, such as click-through rates, to compare treatments. However, these metrics may not align with long-term profit-maximization objectives. We introduce *Bayesian meta-analytic methods* for estimating an overall evaluation criterion that reflects long-term experimental impacts by pooling information across related A/B tests using experiment-level summary statistics. Operationally, the methods place a shared prior on vectors containing both short-term and long-term causal effects, and they update this prior with effect estimates and covariance matrices. The resulting posterior learns cross-metric relationships from historical experiments, allowing currently observed short-term effects to inform long-term business impacts without requiring customer-level data. We compare these approaches via simulation and in an empirical application involving 190 online A/B tests from Zillow using Gaussian, Student-*t*, and Dirichlet process mixture priors. Compared with a single-test benchmark, our best-performing method reduces the posterior standard deviations by about 60%; similarly, a sample-split comparison shows a reduction of 25% in absolute prediction error while maintaining near-nominal predictive coverage. The framework also mitigates overestimation caused by selecting variants with large positive estimates, and it provides insight into latent effect heterogeneity and clustering structure among A/B tests.

History: Hema Yoganarasimhan served as the senior editor.
Funding: This study was funded by Zillow Group.
Supplemental Material: The data files are available at https://doi.org/10.1287/mksc.2025.0164.

Keywords: advertising measurement • digital marketing • experimentation platform • meta-analysis • online experiment • product development

1. Introduction

A/B testing has become an essential tool in digital marketing, providing a systematic approach to assess interventions—web-page layouts, promotional inducements, and product recommendations among others—based on short-term observable outcomes, like click-through or conversion rates. These metrics offer a convenient, rapid way to compare treatments and enact decisions. However, reliance on such measures can misalign with broader downstream business objectives, like sustained revenue growth, customer retention, and brand loyalty, especially in industries with long consideration cycles, such as real estate. This incongruity highlights a methodological and practical gap; even unbiased, low-variance measures of short-term success may not correspond to those indicating sustained impact, so prioritizing them may lead firms astray.

In practice, product and marketing teams must act before downstream outcomes fully accrue, and short-run proxies seldom satisfy the conditions needed to stand in for long-run objectives. Practitioners typically track *multiple* upstream metrics whose signals can conflict—some improve, whereas others worsen—producing indeterminate patterns that complicate decision making. For example, when historically similar tests indicate that initial gains on a short-run metric (e.g., click-through) tend to coincide with downstream losses (e.g., six-month retention), treating the short-run uplift as a "win" can erode value months later. A related situation can arise with loyalty programs; sign-ups may surge initially, yet long-run engagement and spending can diverge from these early signals. In view of these challenges, industry professionals emphasize the importance of developing an overall evaluation

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

criterion (OEC): a high-precision metric based on short-term indicators that produces accurate estimates of long-term business value (Kohavi et al. 2020, chapter 7). Commonly, the OEC is specified as a single metric or a weighted combination of metrics—often "by committee" rather than through analytical means—and consequently, may not produce reliable forecasts of long-term value.

To address this gap, we propose a meta-analytic framework that learns a mapping from short-term effect estimates to long-term business value. In our framework, analysts specify a target OEC as a function of the true (but unknown) long-term causal effects for a given test. Then, using estimates of both short-term and long-term causal effects from previous experiments, the framework learns the cross-experiment relationship between them and produces OEC estimates for the current test based on its available short-term effect estimates. Methodologically, this is accomplished by (i) specifying a shared prior distribution for the full causal-effect vector across a database of historical A/B tests and (ii) applying Bayesian meta-analytic methods to learn the parameters of this distribution, including the associations between short-term and long-term effects.

We show that this approach produces early, uncertainty-quantified inference for the OEC without imposing strong causal assumptions, such as the existence of a surrogate outcome that fully mediates effects of the treatment on the primary outcome. In effect, our approach replaces these causal assumptions with the statistical assumption of a shared experiment-level effect distribution. This assumption is especially credible in settings with many distinct and closely related experiments, such as experiments within a given platform, application, or website. However, it is less plausible when experiments overlap in time or derive from different domains or customer segments. When this assumption holds, we can learn the empirical strength of surrogacy relationships among causal effects in the form of cross-experiment correlations.

Because the approach is fully Bayesian and operates on experiment-level summaries (effect estimates and covariance matrices), it can integrate naturally with complementary measurement systems (e.g., multi-touch attribution) without requiring customer-level data (Berman 2018, Yang et al. 2024). Unlike surrogate index approaches that rely on individual-level surrogacy and comparability assumptions, information transfer in our framework is governed by a shared distribution for experiment-level causal effects. Under our assumption of a shared effect distribution, we develop empirical surrogacy diagnostics (variance-ratio decompositions and cross-trial R^2) that quantify when short-term proxy metrics are informative for long-term metrics and conversely, signal when transfer should be limited. When historically similar tests indicate that initial gains on a short-run metric tend to coincide with downstream losses, our framework regularizes the current experiment's estimates toward that empirically learned relationship. Conversely, when the diagnostics indicate weak or unstable associations, the framework tempers transfer, protecting against overconfident extrapolation.

In short, our primary contribution is to demonstrate how meta-analysis can be used to construct an OEC that links short-term proxies to long-term marketing outcomes using experiment-level summaries. Secondary contributions include empirical surrogacy diagnostics, an empirical assessment across a large archive of industry experiments, and open-source implementation materials to support reproducibility and extension. We illustrate the framework with three random-effects specifications—multivariate Gaussian (G), multivariate Student-t (T), and a Dirichlet process mixture (DPM)—to assess robustness to heavy tails and latent clustering. In simulation and in an application to 190 A/B tests from Zillow, the meta-analytic methods yield materially better inference than single-test benchmarks; the best sample-split method reduces log absolute prediction error by 0.29 points, corresponding to about 25% lower absolute prediction error on the original scale, and posterior standard deviations are about 60% smaller without compromising nominal coverage of prediction intervals. The framework also reduces overstatement when tests are reported or implemented after selecting variants with large positive estimated effects and produces actionable insights related to cross-experiment effect heterogeneity (e.g., posterior clustering under the DPM).

The remainder of the paper proceeds as follows. Section 2 reviews closely related work on long-term inference, surrogate modeling, and information transfer across experiments. Section 3 introduces the OEC meta-analytic framework, and Section 4 details the three model instantiations. Section 5 reports a simulation study comparing the methods on a range of accuracy measures (e.g., credible interval coverage). Section 6 applies the methods to 190 A/B tests from Zillow, including surrogacy diagnostics and an assessment of cross-experiment transfer. Section 7 concludes with a summary, managerial implications, and potential extensions of the framework.

2. Literature Review

The landscape of marketing experimentation has evolved substantially over the past decade, with growing reliance on A/B testing and related designs to guide product and marketing decisions. Despite these advances, existing approaches often fall short when the goal is to *aggregate* evidence across experiments to learn how short-run proxies map to long-run business outcomes. In what follows, we focus on literatures that

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

bear directly on this problem: (i) research on short-term and long-term outcomes in marketing, (ii) statistical methods for meta-analysis and cross-experiment learning, and (iii) Bayesian nonparametric methods for robust inference in marketing applications. For brevity, we eschew a full review of broader topics in A/B practice, adaptive experimentation, and general Bayesian nonparametrics, except where they are directly pertinent to these aims.

2.1. Short-Term Metrics in Marketing

Many marketing decisions are evaluated on the basis of near-term quantities—clicks, sign-ups, and flow completions—that are timely and scalable but may be imperfect surrogates for long-run objectives, such as revenue, retention, and customer equity. This tension appears across common experimentation settings, where immediate responses can obscure downstream effects that accrue over months or longer. In advertising, for example, success is frequently gauged by short-run responses (impressions, clicks, and conversions), yet these signals need not reflect durable acquisition or customer lifetime value; evaluating return on investment, therefore, requires attention to longer-horizon outcomes beyond immediate response (Srinivasan et al. 2010, Lewis and Rao 2015). Likewise, in customer management, interventions that lift early engagement (e.g., loyalty program sign-ups) can diverge from long-run participation and spending patterns unless sustained incentives are in place (Ascarza 2018). These cases illustrate a general challenge; short-term proxies are informative but incomplete, and their link to long-run outcomes is ultimately empirical rather than guaranteed. Because such patterns recur in digital product and marketing contexts, methods that formally link short-run metrics to long-run objectives are particularly valuable.

2.2. Cross-Experiment Learning

The basic meta-analysis paradigm that we follow has been applied extensively in the analysis of clinical trials, enabling the use of "surrogate end points," such as blood pressure, to stand in as indicators of long-term health outcomes, such as five-year mortality, thereby accelerating the decision-making process. In addition to the meta-analysis paradigm, this literature includes the causal effectiveness (CE) and causal association (CA) paradigms, entailing mediation analysis and principal stratification, respectively; both require restrictive causal assumptions for identification (Joffe and Greene 2009, Conlon et al. 2017). In contrast, meta-analytic methods do not require such assumptions, relying instead on the ability to learn effect distributions across historical experiments (Elliott 2023). Whereas clinical applications are often constrained by the scarcity of trials, A/B testing routinely yields extensive archives, rendering the meta-analytic paradigm a natural fit for OEC construction.

The standard approach for statistical meta-analysis is to specify a random-effects model with a shared fixed effect and random effects representing the true but unknown causal effects for each experiment (DerSimonian and Laird 2015). DerSimonian and Laird (1986) originally proposed this approach in a univariate setting, and it was subsequently extended to multivariate random-effects meta-analysis (MREMA), enabling its use in modeling surrogate end points (Buyse et al. 2000). To improve the robustness of the meta-analysis approach, researchers have proposed flexible models for the effect distribution, including the DPM approach that we adopt in Section 4.3 (Burr and Doss 2005). MREMA models can be used to target a wide variety of estimands, including trial-specific effects, average cross-trial parameters, and measures of causal surrogacy (Higgins et al. 2009). They have also been shown to avoid "surrogacy paradoxes," in which other methods (such as the CA and CE frameworks) may produce incorrect inferences (Elliott et al. 2015). In summary, this literature suggests that MREMA methods provide a well-suited framework for long-term inference in A/B testing, especially when coupled with flexible, nonparametric modeling and surrogacy measures.

Recently, there has been increased interest in data combination methods and inference for long-term outcomes in other areas of science and business. One representative approach is that of "surrogate indices" in economics (Imbens et al. 2024, Athey et al. 2026). Yang et al. (2024) adapt this approach to learn targeting policies, enabling them to learn optimal individual-level treatment rules under surrogacy[1] and comparability assumptions. These approaches are best suited to policy-learning settings with individual-level data and identifying assumptions linking short-term and long-term potential outcomes across data sets. In contrast, our focus is experiment-level OEC estimation with calibrated uncertainty from *summary* data (effect estimates and within-experiment covariances), where information transfer is governed by exchangeability across experiments and learned correlations.

A different approach exploits cross-study structure using matrix-completion ideas to recover missing outcomes across studies (Rafieian 2023), and it is conceptually aligned with the broader transfer-learning perspective that leverages stable structure across related tasks to enable information transfer (Pan and Yang 2009). These methods are conceptually similar to our MREMA approach in that they borrow information across data sets, but they generally require individual-level data, introduce more restrictive assumptions, and/or target a different problem, such as policy learning.

2.3. Bayesian Nonparametrics and OEC Construction

Within marketing, Bayesian nonparametric models offer flexible, data-adaptive ways to capture heterogeneity

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

and clustering without prespecifying segment structure. Representative applications include DPMs for random-coefficient heterogeneity in discrete choice (Kim et al. 2004), semiparametric Dirichlet process (DP)-based e-customization that accommodates complex multisource variation (Ansari and Mela 2003), and Gaussian processes for customer-base dynamics that yield interpretable and decision-oriented summaries (Dew and Ansari 2018). Collectively, these approaches suggest the power and generality of using flexible priors to discover latent structure and provide robustness in aggregation, principles especially relevant when pooling effects across experiments and summarizing multiple short-run signals into an OEC.

In summary, recent work has begun to address the business challenges of complex customer heterogeneity and long-term inference/targeting. Meanwhile, previous work in biostatistics has developed the MREMA framework for addressing closely related challenges in clinical trials, enabling robust inference for effects on long-term, noisy outcomes via short-term surrogate metrics. However, the MREMA approach has not yet been applied in A/B testing to (i) aggregate evidence across many experiments and metrics; (ii) deliver early, uncertainty-calibrated OEC forecasts; (iii) quantify surrogacy relationships among related business metrics; and (iv) explicitly account for test-level heterogeneity. The aim of this work is to improve A/B test decision-making practices by filling this gap.

3. Proposed Framework

This section outlines our proposed meta-analysis framework for OEC estimation.

3.1. Basic Setup

We assume access to $T \in \mathbb{N}$ historical A/B tests (experiments). To simplify the exposition, we assume that each test has two arms. The extension to three or more arms is conceptually straightforward; see Appendix A. In each test indexed by $t \in [T] := \{1, \ldots, T\}$, we observe a vector, $\mathbf{z}_t \in \mathbb{R}^p$, of estimated treatment effects for p business metrics. The estimate $\mathbf{z}_t$ is formed from a sample of $n_t \in \mathbb{N}$ observations. Typically, each observation represents a single customer. Our methods are agnostic to the specific estimator defining $\mathbf{z}_t$ provided that the following assumption holds.

Assumption 1 (Asymptotic Normality). *As $n_t \to \infty$, the limiting distribution of $\mathbf{z}_t$ is multivariate Gaussian with mean $\boldsymbol{\mu}_t \in \mathbb{R}^p$ and positive definite covariance matrix $\boldsymbol{\Sigma}_t \in \mathbb{R}^{p \times p}$.*

Assumption 1 ensures that the distribution of $\mathbf{z}_t$ is well approximated by a multivariate Gaussian distribution for large n_t. In practice, n_t must be large enough that the multivariate Gaussian distribution can serve as an accurate approximate likelihood function for $\boldsymbol{\mu}_t$ (van Houwelingen et al. 2002). As usual, precise estimators (those with small $\boldsymbol{\Sigma}_t$) are preferred, and efficiency gains in $\mathbf{z}_t$ directly translate into efficiency gains within the meta-analysis framework. In particular, analysts can often reduce variances by 50% or more by leveraging pretreatment covariates for individual customers (Deng et al. 2013, Jin and Ba 2023).

In practice, we often have access only to an estimator, $\widehat{\boldsymbol{\Sigma}}_t$, of $\boldsymbol{\Sigma}_t$ that satisfies the following assumption.

Assumption 2 (Consistency). *The estimation error $n_t(\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}_t)$ converges in probability to the zero matrix $\mathbf{0}_{p,p}$ as $n_t \to \infty$.*

In our notation, the dependence on n_t is already built into $\widehat{\boldsymbol{\Sigma}}_t$ and $\boldsymbol{\Sigma}_t$; thus, the n_t in Assumption 2 serves only to place these matrices on a constant scale. By Slutsky's theorem, the estimation error is asymptotically negligible, so we take the common meta-analytic approach of treating $\boldsymbol{\Sigma}_t$ as known (DerSimonian and Laird 1986). Consequently, in the remainder of the paper, we remove the hat, denoting $\widehat{\boldsymbol{\Sigma}}_t$ as $\boldsymbol{\Sigma}_t$.

In common practice, firms analyze one metric at a time, which requires only the diagonal elements of $\boldsymbol{\Sigma}_t$. Thus, our framework entails the additional practical requirements of (i) computing metric covariances and (ii) maintaining an archive of comparable experimental results. In particular, the latter requires shared metric definitions across experiments. Crucially, however, our framework requires only these summary statistics, $(\mathbf{z}_t, \boldsymbol{\Sigma}_t)_{t \in [T]}$, removing the need for customer-level data access and processing. In our empirical study, this requirement reduces the size of the data set from billions to hundreds of rows.

We partition $\mathbf{z}_t$ into two sets of metrics: $\mathbf{z}_t = (\mathbf{y}_t^\top, \mathbf{x}_t^\top)^\top$ with $\mathbf{y}_t \in \mathbb{R}^r$ and $\mathbf{x}_t \in \mathbb{R}^q$. The true latent effects, $\boldsymbol{\mu}_t$, can then be partitioned in a corresponding fashion as $\boldsymbol{\mu}_t = (\boldsymbol{\mu}_{\mathbf{y}t}^\top, \boldsymbol{\mu}_{\mathbf{x}t}^\top)^\top$. The vectors $\mathbf{y}_t$ and $\boldsymbol{\mu}_{\mathbf{y}t}$ represent *key performance indicators (KPIs)* for the business units affected by the experiments. The OEC is defined as some analyst-specified function $f : \mathbb{R}^r \to \mathbb{R}$ of the true effects for these KPIs: $\text{OEC}_t := f(\boldsymbol{\mu}_{\mathbf{y}t})$. Analysts should specify $\mathbf{y}_t$ and f such that the OEC represents long-term business value in the most objective manner possible. Typically, the OEC will attempt to quantify customer-level profit or revenue subject to technical constraints and the requirement that $\mathbf{y}_t$ is observed for a large subset of previous experiments.

The remaining metrics comprising $\mathbf{x}_t$ are *surrogates* intended to improve our estimates of $\boldsymbol{\mu}_{\mathbf{y}t}$ and often, are not of direct interest. The surrogates may be measured over a shorter time frame than the KPIs, but this is not a requirement. Both $\mathbf{x}_t$ and $\mathbf{y}_t$ represent estimated effects on post-treatment outcomes; neither is a covariate vector. The definitions of $\mathbf{x}_t$ and $\mathbf{y}_t$ should be standardized across all T experiments. Ideally, they should also include a specific time frame to improve comparability across experiments. For concreteness, we provide several example specifications below.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Example 1 (Total Revenue). We set $r = 1$ and specify μ_{yt} as the average treatment effect (ATE) on revenue over a period of six months (i.e., $\mu_{yt} := \mathbb{E}[m_t(1) - m_t(0)]$, where $m_t(a)$ is the potential outcome (Rubin 1974) of the business metric, m, under treatment $a \in \{0,1\}$ in the test of interest).[2] In this case, f is simply the identity function: $f(\mu_{yt}) = \mu_{yt}$. We would then specify $\mathbf{x}_t$ as a vector of treatment-effect estimates that act as (imperfect) surrogates or proxies for μ_{yt}, such as click-through rates, average order sizes, or measures of customer engagement.

Example 2 (Segmented Revenue). We allow $r > 1$ and specify $\boldsymbol{\mu}_{\mathbf{y}t}$ as a vector of treatment effects with each entry corresponding to an ATE for revenue within a specific product line. In this case, we specify f as $f(\boldsymbol{\mu}_{\mathbf{y}t}) = \mathbf{1}_r^\top \boldsymbol{\mu}_{\mathbf{y}t}$, where $\mathbf{1}_r$ is an r-dimensional vector of ones (i.e., we sum over product lines). We would again specify $\mathbf{x}_t$ as a vector of proxy treatment-effect estimates.

Example 3 (Conversion Funnel). We assume that the metrics comprise a funnel with r layers and specify each component of $\boldsymbol{\mu}_{\mathbf{y}t}$ as a percentage change (lift) at a particular point in the funnel; the jth element of $\boldsymbol{\mu}_{\mathbf{y}t}$ is

$$\mu_{ytj} := \frac{\mathbb{E}[m_{tj}(1) \mid m_{t,j-1}(1) = 1]}{\mathbb{E}[m_{tj}(0) \mid m_{t,j-1}(0) = 1]} - 1,$$

where $m_{tj}(a) \in \{0,1\}$ is the potential outcome for conversion in the jth funnel layer and $m_{t0}(a) = 1$ for $a \in \{0,1\}$. We would then specify $\mathbf{x}_t$ as a vector of proxies for these conversion rates and specify f as the bottom-of-funnel lift:

$$f(\boldsymbol{\mu}_{\mathbf{y}t}) := \prod_{j=1}^{r}(1 + \mu_{ytj}) - 1.$$

Example 3 reveals that the framework is not restricted to ATEs. In practice, analysts may also transform $\boldsymbol{\mu}_t$ and $\mathbf{z}_t$; after applying the multivariate delta method to obtain the corresponding $\boldsymbol{\Sigma}_t$, Assumptions 1 and 2 apply to these transformed quantities. This approach allows us to satisfy Assumptions 1 and 2 while maintaining the natural business scale of the OEC. In the case of Example 3, one such transformation is

$$\mu_{ytj} = \log\left(\frac{\mathbb{E}[m_{tj}(1) \mid m_{t,j-1}(1) = 1]}{\mathbb{E}[m_{tj}(0) \mid m_{t,j-1}(0) = 1]}\right).$$

The OEC_t can then be expressed as $f(\boldsymbol{\mu}_{\mathbf{y}t}) := \exp(\mathbf{1}_r^\top \boldsymbol{\mu}_{\mathbf{y}t}) - 1$. Example 3 also provides an example in which the framework may provide benefit even when metrics are measured over the same time period. Although the bottom-of-funnel events provide the most direct business value, they are typically quite rare, resulting in low-precision estimates relative to the upper-funnel metrics. In this case, the meta-analysis framework enables us to define success in terms of bottom-of-funnel metrics but reap the statistical benefits of using sensitive top-of-funnel metrics.

3.2. The Naive Method

Given the setup of Section 3.1, the problem reduces to that of obtaining inferences for $\mathrm{OEC}_t = f(\boldsymbol{\mu}_{\mathbf{y}t})$ given $\mathbf{z}_t$ and $\boldsymbol{\Sigma}_t$. As a starting point, we consider the following plug-in estimator:

$$\widehat{\mathrm{OEC}}_t := f(\mathbf{y}_t). \tag{1}$$

We refer to $\widehat{\mathrm{OEC}}_t$ as the naive (N) method because it ignores $\mathbf{x}_t$ and data from other experiments. Below, we let $\boldsymbol{\Sigma}_{\mathbf{yy}t}$ be the submatrix of $\boldsymbol{\Sigma}_t$ corresponding to $\mathbf{y}_t$. Under the following assumption, $\widehat{\mathrm{OEC}}_t$ has a simple asymptotic distribution.

Assumption 3 (Regularity). *The function f has continuous partial derivatives at the point $\boldsymbol{\mu}_{\mathbf{y}t}$ and $f'(\boldsymbol{\mu}_{\mathbf{y}t})^\top \boldsymbol{\Sigma}_{\mathbf{yy}t} f'(\boldsymbol{\mu}_{\mathbf{y}t}) > 0$.*

Under Assumption 3, the multivariate delta method gives

$$\frac{\widehat{\mathrm{OEC}}_t - \mathrm{OEC}_t}{\sqrt{f'(\boldsymbol{\mu}_{\mathbf{y}t})^\top \boldsymbol{\Sigma}_{\mathbf{yy}t} f'(\boldsymbol{\mu}_{\mathbf{y}t})}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n_t \to \infty, \tag{2}$$

where $\mathcal{N}$ denotes the Gaussian distribution and $f'(\boldsymbol{\mu}_{\mathbf{y}t})$ denotes the gradient of f. In particular, when $f(\boldsymbol{\mu}_{\mathbf{y}t}) = \mathbf{1}_r^\top \boldsymbol{\mu}_{\mathbf{y}t}$ (Example 2), the asymptotic variance is $\mathbf{1}_r^\top \boldsymbol{\Sigma}_{\mathbf{yy}t} \mathbf{1}_r$, where $\mathbf{1}_r$ is an r-vector of ones. See Deng and Shi (2016) for a related usage of the multivariate delta method in online experimentation.

In Sections 5 and 6, we employ the naive method as a benchmark because it mimics the common industry practice of forming plug-in estimators using data from a single A/B test. Commonly, practitioners specify $\mathbf{z}_t$ as the difference-in-means estimator: $\mathbf{z}_t = \bar{\mathbf{m}}_{t1} - \bar{\mathbf{m}}_{t0}$, where $\bar{\mathbf{m}}_{t1}$ and $\bar{\mathbf{m}}_{t0}$ denote the sample means of the metrics within the treatment and control groups, respectively. The covariance matrix, $\boldsymbol{\Sigma}_t$, can then be estimated as $\mathbf{S}_{t1}/n_{t1} + \mathbf{S}_{t0}/n_{t0}$, where $\mathbf{S}_{t1}$, $\mathbf{S}_{t0}$ and n_{t1}, n_{t0} are the sample covariance matrices and sample sizes in the treatment and control groups, respectively.[3]

In our analysis, we employ these standard estimators in our specification of $\mathbf{z}_t$ and $\boldsymbol{\Sigma}_t$. We reiterate, however, that any estimator satisfying Assumptions 1 and 2 is permitted. If observation-level covariate information is available, practitioners may instead apply advanced causal inference methods, such as causal machine learning, to improve the statistical efficiency of $\mathbf{z}_t$ (Deng et al. 2013, Hair and Sarstedt 2021, Feuerriegel et al. 2024). Although these and other "supervised learning" methods can (and should) be used for generating $\mathbf{z}_t$ where possible, they focus only on *within-experiment* learning. In contrast, our proposed framework targets the largely

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

orthogonal problem of *between-experiment* learning; given a set of estimates $(\mathbf{z}_t)_{t\in[T]}$ and accompanying measurement-error covariances $(\boldsymbol{\Sigma}_t)_{t\in[T]}$, we aim to improve these estimates by borrowing information across experiments. In other words, we aim to develop an estimator $\widetilde{\text{OEC}}_t$—dependent on $(\mathbf{z}_s, \boldsymbol{\Sigma}_s)_{s\in[T]}$—that in some meaningful sense, outperforms $\widehat{\text{OEC}}_t$ as an estimator of OEC_t. The next section provides the basic framework for producing such an estimator.

3.3. The Bayesian Meta-Analysis Framework

To motivate our proposed framework, let $\ell : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be a loss function with $\ell(\widetilde{\text{OEC}}_t, \text{OEC}_t)$ denoting the realized loss for $\widetilde{\text{OEC}}_t$ as an estimate of OEC_t, such as the squared-error loss, $\ell(\widetilde{\text{OEC}}_t, \text{OEC}_t) = (\widetilde{\text{OEC}}_t - \text{OEC}_t)^2$. To define a suitably averaged loss, we introduce the following assumption.

Assumption 4 *Let $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ be a family of distributions with parameter space $\boldsymbol{\Theta}$. For some $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, the true effects $(\boldsymbol{\mu}_t)_{t\in[T]}$ are independent and identically distributed (i.i.d.) according to $P_{\boldsymbol{\theta}}$: $\boldsymbol{\mu}_t \overset{iid}{\sim} P_{\boldsymbol{\theta}}$.*

Assumption 4 is a standard assumption in the meta-analysis literature (Elliott et al. 2015, Elliott 2023), and univariate versions of Assumption 4 have been employed in previous work on online experimentation (Deng 2015; Deng and Shi 2016; Azevedo et al. 2019, 2020; Tripuraneni et al. 2024). Assumption 4 requires that the effects for all tests arise from a common distribution, $P_{\boldsymbol{\theta}}$, given $(\boldsymbol{\Sigma}_t)_{t\in[T]}$ and that these effects are independent, ruling out (for example) cross-test interference. In practice, the latter can be enforced through thoughtful experimental design, such as independently randomizing arm assignments and sequencing A/B tests that may interfere with each other.

Assumption 4 is the key assumption within our meta-analytic framework. Relative to the CA, CE, and surrogate index approaches, the key requirement is different; instead of imposing additional causal surrogacy assumptions linking short-term and long-term potential outcomes, we assume that the vector of experiment-level causal effects arises from a shared distribution across tests. This assumption will be violated when the assumed model family $P_{\boldsymbol{\theta}}$ is not large enough to contain the true effect distribution, but subtler violations are possible even with a nonparametric model class, such as violations of the requirement that the experiments arise from the same distribution. As such, Assumption 4 is most appropriate in settings with homogeneous experiments, arising from a single domain and population. Assumption 4 represents somewhat of an idealized setting; however, we show below that it leads to estimators with an approximate optimality property.

Assumption 4 imposes observable consequences, so we may attempt to falsify it using posterior predictive model checks (Rubin 1984, Meng 1994, Gelman et al. 1996). In particular, we recommend checking for dependence between $\boldsymbol{\mu}_t$ and known experiment-level metadata, including $\boldsymbol{\Sigma}_t$, as this dependence would signal departures from the assumption of identical distributions. If analysts do find significant dependence, we recommend adopting the metaregression (MR) approach discussed in Section 3.4. This approach allows us to weaken Assumption 4, requiring only conditional independence given $\boldsymbol{\Sigma}_t$ and potentially other experiment-level metadata.

Because $\boldsymbol{\mu}_t$ is vector valued, $P_{\boldsymbol{\theta}}$ encodes information regarding the metrics' location, scale, and dependence structure. Under departures from independence, the latter enables the metrics to act as surrogates for one another. Moreover, Assumption 4 allows us to define a notion of optimality in a setting that acknowledges the randomness of $(\boldsymbol{\mu}_t)_{t\in[T]}$ as opposed to treating all $(\boldsymbol{\mu}_t)_{t\in[T]}$ as unrelated. Under Assumption 4, the *Bayes risk* of the estimator $\widetilde{\text{OEC}}_t$ is defined as $\mathbb{E}_{\boldsymbol{\theta}}\,\ell(\widetilde{\text{OEC}}_t, \text{OEC}_t)$ with the expectation taken over both $\mathbf{z}_t$ and $\boldsymbol{\mu}_t$, the latter being with respect to $P_{\boldsymbol{\theta}}$. Under weak regularity conditions (see Brown and Purves 1973 and Berger 1985, section 4.4.1), a minimizer of the Bayes risk can be computed by minimizing the posterior risk given $\mathbf{z}_t$:

$$\widetilde{\text{OEC}}_{\boldsymbol{\theta}t} := \arg\min_{g} \mathbb{E}_{\boldsymbol{\theta}}[\ell\{g(\mathbf{z}_t, \boldsymbol{\Sigma}_t), \text{OEC}_t\} \mid \mathbf{z}_t]. \tag{3}$$

In particular, under squared-error loss, it is well known that the minimizer of (3) is the posterior mean of $f(\boldsymbol{\mu}_{\mathbf{y}t})$. Equation (3) omits $\boldsymbol{\Sigma}_t$ from the conditioning set because we treat it as a fixed, nonrandom quantity throughout.

In practice, $\widetilde{\text{OEC}}_{\boldsymbol{\theta}t}$ is not a feasible estimator because $\boldsymbol{\theta}$ is unknown, but it motivates estimators that match its logic of minimizing posterior risk given an estimate of $\boldsymbol{\theta}$. We adopt a fully Bayesian approach, placing a hyperprior[4] on $\boldsymbol{\theta}$ with hyperparameter $\boldsymbol{\theta}_0$. As with $\boldsymbol{\theta}$, we do not impose any specific parametric structure or dimension on $\boldsymbol{\theta}_0$ at this stage. For test t, our analysis involves summarizing the posterior distribution with density $p\{\text{OEC}_t \mid (\mathbf{z}_s)_{s\in[T]}\}$ proportional to

$$\int \delta\{\text{OEC}_t - f(\boldsymbol{\mu}_{\mathbf{y}t})\}\, p(\boldsymbol{\theta}\mid\boldsymbol{\theta}_0)\cdot\prod_{s=1}^{T} p(\boldsymbol{\mu}_s\mid\boldsymbol{\theta})\, p(\mathbf{z}_s\mid\boldsymbol{\mu}_s)\, d\boldsymbol{\zeta}, \tag{4}$$

where $p(\cdot\mid\cdot)$ denotes the conditional density of its arguments, $\delta(\cdot)$ is the Dirac delta function, and $\boldsymbol{\zeta} := \{\boldsymbol{\theta}, (\boldsymbol{\mu}_s)_{s\in[T]}\}$. The integral in (4) marginalizes over $\boldsymbol{\theta}$ and the components of $(\boldsymbol{\mu}_s)_{s\in[T]}$ not equal to $\boldsymbol{\mu}_{\mathbf{y}t}$. As above, we omit $(\boldsymbol{\Sigma}_s)_{s\in[T]}$ from the conditioning sets because we treat it as a fixed quantity.

In practice, the integral in (4) can be approximated using posterior draws from a Markov chain Monte

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Carlo (MCMC) algorithm. Concretely, this is accomplished by applying f to posterior draws of $\boldsymbol{\mu}_{\mathbf{y}t}$. Given a particular loss function, we can produce estimators minimizing the posterior loss in an analogous fashion to (3), replacing the unknown posterior with (4). Although the resulting estimators are not Bayes-risk optimal in finite samples, they are optimal under an enlarged model that includes $\boldsymbol{\theta}_0$ (i.e., they are optimal if $\boldsymbol{\theta}$ is sampled according to the hyperprior density, $p(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$). Under misspecification of the hyperprior, Bayesian methods—including nonparametric models, like the DPM—can still achieve posterior consistency, meaning that the posterior distribution concentrates around the correct value of $\boldsymbol{\theta}$ (Rousseau 2016, Ray and van der Vaart 2021). Consequently, we expect our proposed meta-analysis methods to closely mimic the optimal estimator, $\widetilde{\mathrm{OEC}}_{\boldsymbol{\theta}t}$, in large samples provided that the relevant regularity conditions are satisfied.

Figure 1 summarizes the generic meta-analysis model. Panel (a) of Figure 1 visualizes the model in the form of a directed acyclic graph (DAG). Panel (b) of Figure 1 provides an example of how Bayes' rule adjusts the naive point estimate, $\mathbf{z}_t$, shrinking it toward a region of high prior probability. In this example in panel (b) of Figure 1, the sampling-based (blue) and posterior (red) confidence/credible regions have identical coverage levels, and both are centered around an unbiased estimate; however, the posterior credible region is much smaller, and the accompanying point estimate is Bayes-risk optimal under squared-error loss. This improvement is possible because the likelihood (blue) region in panel (b) of Figure 1 is highly unlikely a priori, so on average, the Bayes estimator will perform much better by adjusting the estimate and accompanying credible region toward the region of high prior probability. Also, notice that μ_{xt} and μ_{yt} exhibit strong positive correlation. Consequently, the model's relatively high confidence that μ_{xt} is large (given the estimate x_t and the sampling variability in panel (b) of Figure 1) causes the model to shift its less certain estimate of μ_{yt} substantially upward.

3.4. Extension to Metaregression

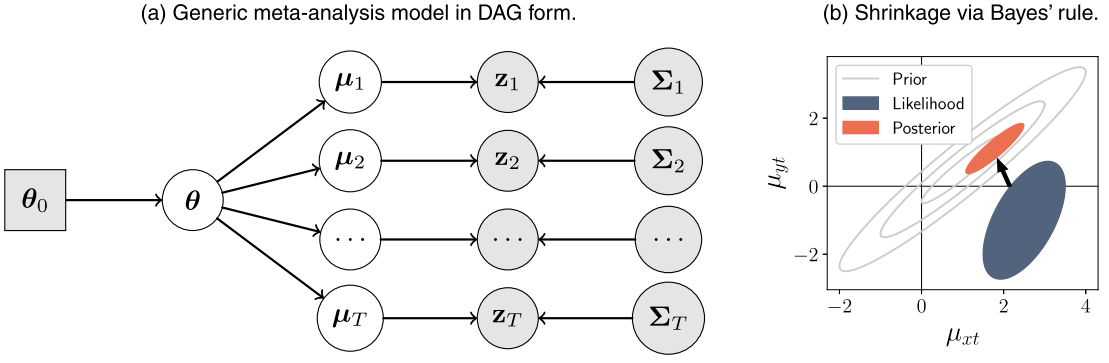
Given experiment-level metadata—target platform, intervention type, test duration, etc.—Assumption 4 can be relaxed to allow the distribution of $\boldsymbol{\mu}_t$ to vary by test, resulting in a metaregression model (van Houwelingen et al. 2002). Mathematically, this information can be encoded in a covariate vector $\mathbf{w}_t \in \mathbb{R}^d$ for some $d \in \mathbb{N}$. Assumption 4 can then be relaxed to require only conditional independence across tests given $\mathbf{w}_t$ and $\boldsymbol{\Sigma}_t$; in the assumption below, $\mathbf{w}$ represents a generic value of $\mathbf{w}_t$ and $\overset{ind}{\sim}$ means "independently distributed as."

Assumption 5 (Conditional Independence). *Let* $\{P_{\boldsymbol{\theta},\mathbf{w}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}, \mathbf{w} \in \mathbb{R}^d\}$ *be a family of conditional distributions given* $\mathbf{w}$ *with parameter space* $\boldsymbol{\Theta}$. *Then,* $\boldsymbol{\mu}_t|(\mathbf{w}_s)_{s\in[T]} \overset{ind}{\sim} P_{\boldsymbol{\theta},\mathbf{w}_t}$ *for some* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

The vector $\mathbf{w}_t$ can also include summaries of $\boldsymbol{\Sigma}_t$; this may be required if especially high- or low-performing tests are stopped quickly, resulting in large variances. Analysts may also include time variables in $\mathbf{w}_t$ to capture changes in the effect distribution over time. In general, we recommend using model checking to determine whether a metaregression model is needed and if so, which variables to include in $\mathbf{w}_t$.

The data in our empirical study do not contain rich experiment-level metadata, and we find limited evidence of dependence between $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$. Consequently, with the exception of the simulation study in Section 5, the remainder of the paper focuses on meta-analytic models under Assumption 4. The next section presents three representative models in this context.[5]

Figure 1. (Color online) The Meta-Analysis Model for OEC



Notes. In the directed acyclic graph (DAG) in panel (a), circles denote random variables, and the box denotes a fixed a priori value. Shading denotes values that are observed or known. We assume $\mathbf{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with $\boldsymbol{\mu}_t \overset{ind}{\sim} P_{\boldsymbol{\theta}}$ for all $t \in [T]$. The vector $\boldsymbol{\theta}_0$ denotes fixed hyperparameters for the prior distribution on $\boldsymbol{\theta}$. The plot in panel (b) provides an example in which Bayesian shrinkage significantly improves the point estimate, $\mathbf{z}_t$, and its variance because of the strong prior correlation between μ_{xt} and μ_{yt}. (a) Generic meta-analysis model in DAG form. (b) Shrinkage via Bayes' rule.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

4. Meta-Analysis Models

Having outlined the basic meta-analysis framework, we now provide three specific meta-analysis models that we compare in Sections 5 and 6.

4.1. Multivariate Gaussian Model

We first consider a multivariate Gaussian model for $P_{\boldsymbol{\theta}}$: $\boldsymbol{\mu}_t \mid \boldsymbol{\theta} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Omega})$, where $\boldsymbol{\eta} \in \mathbb{R}^p$ and $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ with $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\Omega})$. Given $\boldsymbol{\eta}$ and $\boldsymbol{\Omega}$ (or posterior samples thereof), it is well known that the posterior distribution of $\boldsymbol{\mu}_t$ is available in closed form:

$$\boldsymbol{\mu}_t \mid \{\mathbf{z}_s, \boldsymbol{\Sigma}_s\}_{s \in [T]}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{N}(\boldsymbol{\eta}_t, \boldsymbol{\Omega}_t), \tag{5a}$$

$$\boldsymbol{\Omega}_t = (\boldsymbol{\Omega}^{-1} + \boldsymbol{\Sigma}_t^{-1})^{-1}, \quad \boldsymbol{\eta}_t = \boldsymbol{\Omega}_t(\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} + \boldsymbol{\Sigma}_t^{-1}\mathbf{z}_t), \tag{5b}$$

where the $\{\boldsymbol{\mu}_t\}_{t \in [T]}$ are independent across t conditional on $(\boldsymbol{\theta}, \{\mathbf{z}_s, \boldsymbol{\Sigma}_s\}_{s \in [T]})$. For additional details, see Rossi et al. (2024, section 2.9.5). In the context of A/B testing, Deng et al. (2023) briefly discuss a special case of this model and show that the addition of the prior leads to a statistical efficiency gain over the naive method. In fact, the posterior mean $\boldsymbol{\eta}_t$ is the best linear unbiased predictor (BLUP) of $\boldsymbol{\mu}_t$ given $\mathbf{z}_t$, $\boldsymbol{\Sigma}_t$, and $\boldsymbol{\theta}$ (Henderson 1950, Robinson 1991). Recent OEC approaches proposed by Google and ShareChat also propose estimators that are linear in $\mathbf{z}_t$ but are not equal to BLUPs (Jeunen and Ustimenko 2024, Tripuraneni et al. 2024). Under the Gaussian random-effects model, the posterior mean is the BLUP and therefore, the optimal linear predictor of $\boldsymbol{\mu}_t$, providing a natural benchmark for other linear OEC estimators.

A particularly desirable feature of the BLUPs is that they represent a linear combination of $\boldsymbol{\eta}$ and $\mathbf{z}_t$, which can aid in interpretation. Further, in a standard asymptotic regime with $\sqrt{n_t}\boldsymbol{\Sigma}_t \xrightarrow{p} \mathbf{0}_{p,p}$, we have $n_t(\boldsymbol{\Omega}_t - \boldsymbol{\Sigma}_t) \xrightarrow{p} \mathbf{0}_{p,p}$ and $\sqrt{n_t}(\boldsymbol{\eta}_t - \mathbf{z}_t) \xrightarrow{p} \mathbf{0}_p$ as $n_t \to \infty$. Thus, the BLUP $\boldsymbol{\eta}_t$ is asymptotically equivalent to the naive estimator, $\mathbf{z}_t$, and produces the same asymptotic inferences. In finite samples, however, the BLUP can substantially outperform the naive estimator, especially when $P_{\boldsymbol{\theta}}$ is highly concentrated—a common scenario in online experimentation.

BLUPs generalize naturally to the setting in which $\mathbf{z}_t$ is only partially observed provided that the missingness is not systematic.[6] In this case, the observed effect estimates can be written in the form $\mathbf{A}\mathbf{z}_t$ for some matrix $\mathbf{A}$; for example, if the first $b \in \mathbb{N}$ elements of $\mathbf{z}_t$ are not observed, we set $\mathbf{A} = [\mathbf{0}_{p-b,b}\, \mathbf{I}_{p-b}]$, where $\mathbf{I}_{p-b}$ is an identity matrix of dimension $p - b$. The analysis is then modified by replacing the term $p(\mathbf{z}_s \mid \boldsymbol{\mu}_s)$ in (4) with $p(\mathbf{A}\mathbf{z}_s \mid \boldsymbol{\mu}_s)$ and modifying the estimation algorithm accordingly. For the Gaussian model, this modification results in the following posterior parameters in place of those given in (5):

$$\boldsymbol{\Omega}_t = [\boldsymbol{\Omega}^{-1} + \mathbf{A}^\top(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top)^{-1}\mathbf{A}]^{-1}, \tag{6a}$$

$$\boldsymbol{\eta}_t = \boldsymbol{\Omega}_t[\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} + \mathbf{A}^\top(\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{z}_t]. \tag{6b}$$

If $\mathbf{y}_t$ is missing, for example, the term $\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top$ is equal to $\boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}t}$: the (invertible) submatrix of $\boldsymbol{\Sigma}_t$ corresponding to $\mathbf{x}_t$. The implication for the OEC meta-analysis framework is that we can obtain inferences for unobserved metrics—including long-term metrics before their time frame has elapsed—provided that we have access to short-term proxies. Notably, this approach does not require restrictive causal assumptions beyond what is needed for standard effect estimation. Rather, it requires that $\boldsymbol{\mu}_{\mathbf{x}t}$ and $\boldsymbol{\mu}_{\mathbf{y}t}$ are correlated; these correlations are encoded in $\boldsymbol{\Omega}$, and stronger correlations result in more precise inference for $\boldsymbol{\mu}_{\mathbf{y}t}$. The Student-$t$ and DPM models can be modified in a similar fashion to allow for missing data.

In practice, $\boldsymbol{\theta}$ is unknown and must be estimated, so posterior summaries for $\boldsymbol{\mu}_t$ and OEC_t effectively average over the BLUPs defined in (5) according to the posterior density associated with different values of $\boldsymbol{\theta}$. Depending on the choice of computational approach, this posterior averaging may occur implicitly. All three models that we consider admit efficient "collapsed" Gibbs samplers (Liu et al. 1994), so we employ this computational approach throughout, although other techniques, such as Hamiltonian Monte Carlo (HMC), could also be applied (Brooks et al. 2011, Neal 2011). The Gibbs sampling methods that we employ are well known (Neal 2000); however, to assist practitioners in applying our methods, we include them in Appendix B with specific suggestions on how to specify the hyperprior. In general, we recommend specifying dispersed, weakly informative prior distributions where possible, allowing the data set to "speak for itself" in terms of specifying plausible values for $\boldsymbol{\theta}$. For small T, this specification has the benefit of producing estimates that closely align with those of the naive method; as T increases, the estimates improve as the model develops a better estimate of $\boldsymbol{\theta}$.

Buyse et al. (2000) were the first to propose a (bivariate) version of this model in the surrogate end-points literature. Buyse et al. (2000) note that the strength of a surrogate can be assessed by the coefficient of determination from a regression of μ_{yt} on μ_{xt}: $R^2 := \Omega_{yx}^2/(\Omega_{xx}\Omega_{yy})$, where the indices denote elements of $\boldsymbol{\Omega}$ corresponding to either μ_{xt} or μ_{yt}. In the case of vector-valued $\boldsymbol{\mu}_{\mathbf{x}t}$, this measure generalizes naturally to $R^2 := \boldsymbol{\Omega}_{\mathbf{y}\mathbf{x}}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}/\Omega_{yy}$. When $\boldsymbol{\mu}_{\mathbf{y}t}$ is also vector valued and we employ the summed OEC of Example 2, we can further generalize this surrogacy measure as

$$R^2 := \frac{\mathbf{1}_r^\top \boldsymbol{\Omega}_{\mathbf{y}\mathbf{x}}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^{-1}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\mathbf{1}_r}{\mathbf{1}_r^\top \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\mathbf{1}_r}, \tag{7}$$

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

which assesses the proportion of variance in OEC_t explained by $\boldsymbol{\mu}_{\mathbf{x}t}$.

For test t, we define the *total variance ratio* (TVR_t) as the ratio of the meta-analysis posterior variance to the naive variance: $\text{TVR}_t := v_{\mathbf{z}t}/v_{0t}$, where $v_{\mathbf{z}t} := \mathbf{u}^\top \boldsymbol{\Omega}_t \mathbf{u}$ and $v_{0t} := \mathbf{u}^\top \boldsymbol{\Sigma}_t \mathbf{u}$ with $\mathbf{u} = (\mathbf{1}_r^\top, \mathbf{0}_q^\top)^\top$. This ratio can be interpreted in terms of cost savings as follows; achieving a similar posterior variance (or standard error) using the naive method would require increasing the sample size by a factor of $1/\text{TVR}_t$. To further understand the source of the variance reduction, we define the intermediate quantity $v_{\mathbf{y}t} := \mathbf{1}_r^\top (\boldsymbol{\Omega}_{\mathbf{yy}}^{-1} + \boldsymbol{\Sigma}_{\mathbf{yy}t}^{-1})^{-1} \mathbf{1}_r$: the posterior variance from a hypothetical meta-analysis using only $\mathbf{y}_t$ (discarding $\mathbf{x}_t$). We can then define the *prior variance ratio* (PVR), $\text{PVR}_t := v_{\mathbf{y}t}/v_{0t}$, and *surrogacy variance ratio* (SVR), $\text{SVR}_t := v_{\mathbf{z}t}/v_{\mathbf{y}t}$, representing the multiplicative effects on the posterior variance because of (i) the prior distribution of $\boldsymbol{\mu}_{\mathbf{y}t}$ and (ii) surrogacy of $\boldsymbol{\mu}_{\mathbf{x}t}$ for $\boldsymbol{\mu}_{\mathbf{y}t}$, respectively. Because the elements of $\boldsymbol{\mu}_{\mathbf{y}t}$ may, in general, correlate with one another, PVR_t includes reduction in the variance attributable to *internal surrogacy*: surrogacy of the elements of $\boldsymbol{\mu}_{\mathbf{y}t}$ for one another.[7]

Having defined these quantities, we can then decompose TVR_t as $\text{TVR}_t = \text{PVR}_t \cdot \text{SVR}_t$ to assess the contribution of each component to the efficiency gain. These metrics can also be computed for the individual components of $\boldsymbol{\mu}_{\mathbf{y}t}$, providing measures of surrogacy for one outcome given all of the others—a strategy that we employ in our analysis in Section 6. In practice, we recommend using these surrogacy metrics to inform which metrics to include in $\mathbf{x}_t$ so as to maximize the strength of surrogacy subject to model complexity constraints.

4.2. Multivariate Student-*t* Model

Previous work has noted that large outliers are empirically much more common in A/B testing than the Gaussian assumption of Section 4.1 would allow (Azevedo et al. 2020). Perhaps the simplest extension of Section 4.1 to capture this phenomenon is to allow $P_{\boldsymbol{\theta}}$ to take the form of a multivariate Student-*t* distribution:

$$\boldsymbol{\mu}_t \mid \boldsymbol{\theta} \overset{iid}{\sim} \mathcal{T}(\boldsymbol{\eta}, \boldsymbol{\Omega}, \gamma), \tag{8}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\Omega}$ maintain their interpretation as location and scale parameters, $\gamma > 0$ is the degrees of freedom parameter, and $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\Omega}, \gamma)$. The expression for R^2 in (7) remains valid provided that $\gamma > 2$ (Cornish 1954, Ding 2016). Because this model is not fully conjugate, closed-form analogs of TVR, PVR, and SVR are not available, and the posterior mean does not possess a simple linear form as in (5b).

4.3. DPM Model

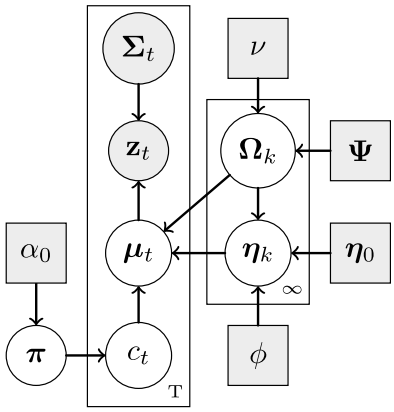
The Gaussian and Student-*t* models serve as a useful starting point, especially when A/B tests are believed to be fairly homogeneous. This section introduces the DPM method, a flexible nonparametric model that avoids placing restrictive parametric assumptions on $P_{\boldsymbol{\theta}}$. The DPM method assumes that $(\boldsymbol{\mu}_t)_{t\in[T]}$ are drawn from a mixture distribution, $P_{\boldsymbol{\theta}}$, where $P_{\boldsymbol{\theta}}$ itself is a random draw from a Dirichlet process with base distribution G_0 and concentration parameter $\alpha_0 > 0$. Formally, the DPM model can be represented as

$$\begin{aligned}
\boldsymbol{\mu}_t \mid c_t, \{\boldsymbol{\eta}_k, \boldsymbol{\Omega}_k\}_{k=1}^{\infty} &\overset{ind}{\sim} \mathcal{N}(\boldsymbol{\eta}_{c_t}, \boldsymbol{\Omega}_{c_t}),\\
c_t \mid \boldsymbol{\pi} &\overset{iid}{\sim} \text{Categorical}(\boldsymbol{\pi}),\\
\boldsymbol{\pi} \mid \alpha_0 &\sim \text{GEM}(\alpha_0),\\
\boldsymbol{\eta}_k, \boldsymbol{\Omega}_k \mid \boldsymbol{\xi} &\overset{iid}{\sim} G_0,
\end{aligned}$$

where c_t denotes the cluster assignment for experiment t; $\boldsymbol{\pi}$ is the vector of cluster probabilities; and GEM denotes the Griffiths, Engen, and McCloskey distribution (Sethuraman and Ghosh 2024). In our assumed model, G_0 is a Gaussian-inverse-Wishart distribution parameterized by $\boldsymbol{\xi} = (\boldsymbol{\Psi}, \nu, \boldsymbol{\eta}_0, \phi)$. Figure 2 summarizes our assumed DPM model in the form of a DAG. The DP prior on $P_{\boldsymbol{\theta}}$ informs a clustering structure among the experiments, whereas the concentration parameter, α_0, dictates the tendency to form new clusters; larger values of α_0 encourage the formation of more clusters.

The surrogacy measures R^2, TVR, PVR, and SVR generalize in a straightforward manner to the DPM model by replacing $\boldsymbol{\Omega}$ with the cluster-specific covariance matrices, $\{\boldsymbol{\Omega}_k\}_{k\in[K]}$, where K is the estimated number of clusters. These measures can be used in a cluster-specific or test-specific manner or aggregated across tests and clusters to obtain integrated measures of surrogacy.

Figure 2. The DPM Model



Notes. Shape type and shading denote whether the variable is random or observed, respectively, as in panel (a) of Figure 1. The grouping rectangles (plates) represent conditionally independent random variables, with the degree shown in the bottom right. In the case of $\boldsymbol{\eta}_k$ and $\boldsymbol{\Omega}_k$, the degree is infinite because the DPM model allows for an infinite number of latent clusters.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

5. Simulation Study

We performed a simulation study to compare the performance of the methods developed in Section 4. The simulations comprise independent replicates with known ground-truth OEC_t values, enabling a richer comparison than is possible in our empirical study in Section 6. In particular, the synthetic nature of the data allows us to compute a larger set of performance metrics and apply a metaregression model based on a synthetic cluster membership variable.

5.1. Simulation Setup

To enforce fidelity between the simulation study and data analysis, we draw 100 independent data sets from the DPM model's posterior predictive distribution after fitting it to the observed data set as described in Section 6.1. However, we apply transformations, including translations, rescalings, and rotations, to ensure that the simulated data set does not reveal any material nonpublic information regarding Zillow's business metrics or experimentation processes. We have made these simulated data sets available on the Open Science Framework to allow researchers to reproduce and build on our results. The simulation analyses use the same methods, prior distributions, and computational algorithms as the empirical analysis described in Section 6.

In addition to testing the three meta-analysis approaches introduced in Section 4, we also implement the naive plug-in method described in (1), which mimics the common industry practice of analyzing A/B tests individually with no shared prior distribution. All methods use the same values of $\mathbf{z}_t$ and $\boldsymbol{\Sigma}_t$ as input. We also include two additional variations of the methods presented in Section 4: Gaussian-0 (G-0) and metaregression. The Gaussian-0 model fixes $\boldsymbol{\eta} = \mathbf{0}$, allowing us to quantify the benefit of estimating the prior metric mean compared with fixing it to zero. MR is a metaregression variant of the DPM model in which the cluster labels are known. From a practical standpoint, this represents a setting in which experiment type is known a priori—for example, mobile versus browser tests—so that the effect distribution may differ across known but related groups. For brevity, the method names are sometimes abbreviated N, G-0, G, T, DPM, and MR.

5.2. Simulation Results

We first compare the methods on six performance metrics.

1. Correct decision rate. The percentage of tests that correctly identify the test arm (A or B) that will maximize OEC_t: B if $OEC_t > 0$ and else, A.

2. Regret. The difference in OEC_t between the selected arm and the optimal arm.

3. Overestimation. The average estimation error (predicted − true) for OEC_t among selected arms. This metric quantifies the overstatement that arises when variants are selected for implementation or reporting because their estimated OEC is positive.

4. Mean absolute error (MAE). The average (across tests) absolute error in the estimation of OEC_t.

5. Average 95% confidence interval (95% CI) length. The average (across tests) length of 95% credible (confidence) intervals for OEC_t.

6. 95% CI coverage. The coverage rate for the above 95% credible (confidence) intervals.

In the results reported below, we first compute these metrics *within* each replication, and then, we compute averages and standard errors *across* replications. Figure 3 displays the average values of the six metrics introduced above with 95% confidence intervals. To enhance the precision of the comparison, we also report the underlying metric values in Table 1.

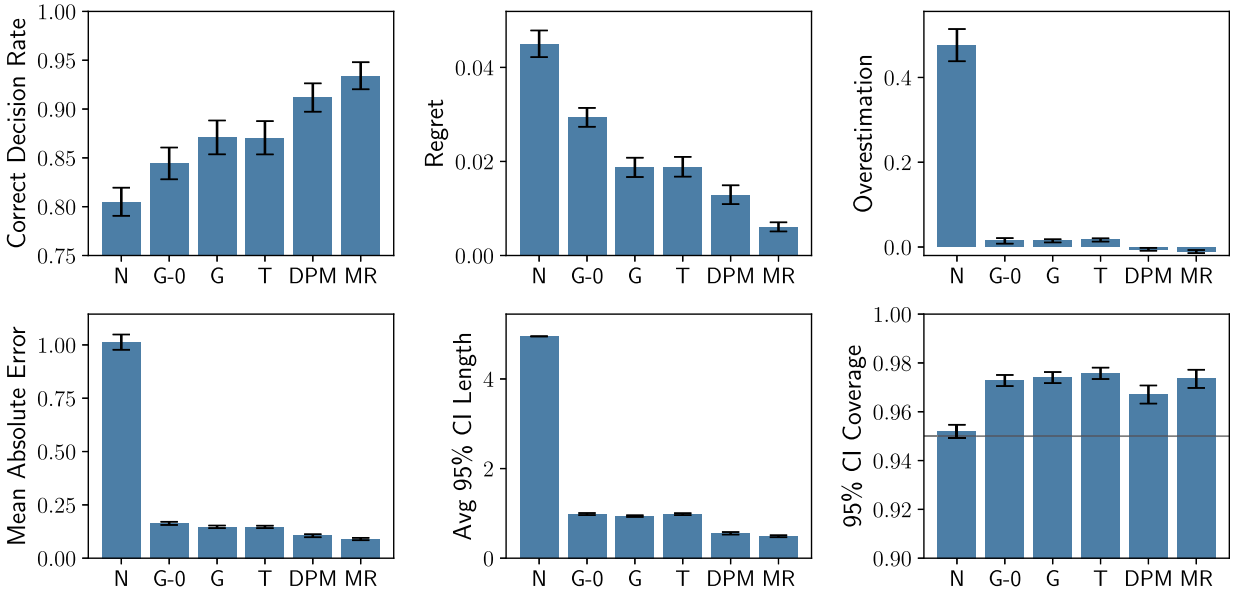
In the upper left panel of Figure 3, we see that the meta-analysis methods correctly identify the winning variant more frequently than the naive method, resulting in lower regret in the upper center panel of Figure 3. In both comparisons, the Gaussian and Student-*t* methods perform similarly, correctly identifying the winning arm in 87% of tests compared with 81% for the naive method, 84% for the Gaussian-0 method, 91% for the DPM method, and 93% for the metaregression method. Thus, compared with the naive method, the DPM and metaregression methods reduce the number of incorrect decisions by more than 50%. The upper right panel of Figure 3 shows that the naive method substantially overestimates the effects of implemented variants, where implementation is based on positive estimated effects, with an average overestimation of 0.48 compared with negligible values in the range of [−0.01 to 0.02] for the meta-analysis methods.

The lower left panel of Figure 3 shows that the MAE follows a pattern similar to that of the above metrics; the metric improves (decreases) from left to right, with the DPM and MR methods performing best. The remaining lower center and lower right panels of Figure 3 compare the 95% credible (confidence) intervals. The meta-analysis methods produce intervals less than one fifth the length of the naive intervals (on average), yet they also cover at a higher rate. As expected, the naive method has near-nominal coverage of 95.2%. In contrast, the meta-analysis methods exhibit mild overcoverage—a consequence of employing misspecified (hyper-)prior distributions.[8] The DPM and MR methods produce the narrowest intervals, presumably because their assumed prior distribution matches the data-generating process, enabling these methods to accurately adjust credible regions based on cluster membership.

Table 2 displays pair-wise comparisons across five[9] of the six metrics listed above and two additional

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Figure 3. (Color online) Simulation Results



Notes. These bar charts visualize the simulation performance detailed in Table 1. The error bars represent 95% confidence intervals. The labels correspond to the naive, Gaussian-0, Gaussian, Student-*t*, Dirichlet process mixture, and metaregression methods, respectively.

metrics: the *Win Rate* and the *Mean* log *Error Quotient* (*MLEQ*). The *Win Rate* is defined as the percentage of A/B tests in which the method listed in the row outperformed the method listed in the column in terms of the magnitude of estimation error. The *MLEQ* is defined as $\log(|a-c|/|b-c|)$, where a is the row prediction, b is the column prediction, and c is the ground-truth value of OEC_t; more negative $\Delta MLEQ$ indicates that the row method has smaller error. The meta-analysis methods outperform the naive method across both the *Win Rate* and *MLEQ* metrics, and the performance ordering is MR, DPM, T, Gaussian, Gaussian-0, and naive from best to worst. The MR and DPM methods outperform each of the other four methods at least 67% of the time (in separate comparisons). The performance order is consistent across the remaining performance metrics (excluding the 95% CI coverage, which is at or above the nominal 95% level for all methods). In total, 98 of 105 pair-wise comparisons are statistically significant, with all 7 nonsignificant comparisons involving the Gaussian-0, Gaussian, and T methods.

In summary, the meta-analysis methods substantially outperform the naive method. The Gaussian and T methods perform comparably with each other and slightly better than the Gaussian-0 method. The DPM and MR methods substantially outperform the other methods. Of these two methods, the MR method performs best, presumably because it has access to the cluster membership variables, which the DPM method must estimate.

6. Application to Real Estate Technology A/B Tests

In this section, we present results from applying the Bayesian meta-analysis methods to 190 A/B tests from Zillow. These tests represent all A/B tests during a subset of 2022–2024 that meet certain data quality standards, such as involving a sufficiently large number of users and not exhibiting unexpected imbalances in the assignment proportions. The A/B tests involve a variety of interventions ranging from new or altered email

Table 1. Simulation Results

	N	G-0	G	T	DPM	MR
Correct Decision Rate	0.81 (0.01)	0.84 (0.01)	0.87 (0.01)	0.87 (0.01)	0.91 (0.01)	0.93 (0.01)
Regret	0.05 (0.00)	0.03 (0.00)	0.02 (0.00)	0.02 (0.00)	0.01 (0.00)	0.01 (0.00)
Overestimation	0.48 (0.02)	0.01 (0.00)	0.01 (0.00)	0.02 (0.00)	−0.01 (0.00)	−0.01 (0.00)
MAE	1.01 (0.02)	0.16 (0.00)	0.15 (0.00)	0.15 (0.00)	0.11 (0.00)	0.09 (0.00)
Avg 95% CI Length	4.95 (0.00)	0.98 (0.01)	0.94 (0.01)	0.98 (0.01)	0.56 (0.01)	0.49 (0.01)
95% CI Coverage	0.95 (0.00)	0.97 (0.00)	0.97 (0.00)	0.98 (0.00)	0.97 (0.00)	0.97 (0.00)

Notes. Performances of the six methods in the simulation study across six performance metrics (standard errors) are shown. The column headings correspond to the naive, Gaussian-0, Gaussian, Student-*t*, Dirichlet process mixture, and metaregression methods, respectively. See Figure 3 for a visualization of these values. Avg, average; MAE, mean absolute error.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Table 2. Simulation Pair-Wise Method Comparisons

Metric	Method	N	G-0	G	T	DPM	MR
Win Rate	N	—	0.25 (0.01)	0.23 (0.01)	0.22 (0.01)	0.21 (0.00)	0.20 (0.00)
	G-0	0.75 (0.01)	—	0.44 (0.01)	0.42 (0.01)	0.30 (0.01)	0.27 (0.00)
	G	0.77 (0.01)	0.56 (0.01)	—	0.45 (0.01)	0.33 (0.01)	0.30 (0.01)
	T	0.78 (0.01)	0.58 (0.01)	0.55 (0.01)	—	0.33 (0.01)	0.30 (0.01)
	DPM	0.79 (0.00)	0.70 (0.01)	0.67 (0.01)	0.67 (0.01)	—	0.42 (0.01)
	MR	0.80 (0.00)	0.73 (0.00)	0.70 (0.01)	0.70 (0.01)	0.58 (0.01)	—
Δ*Mean* log *Error Quotient*	N	—	0.63 (0.02)	0.71 (0.02)	0.72 (0.02)	1.12 (0.03)	1.23 (0.03)
	G-0	−0.63 (0.02)	—	0.08 (0.01)	0.09 (0.01)	0.50 (0.02)	0.60 (0.02)
	G	−0.71 (0.02)	−0.08 (0.01)	—	0.01 (0.01)	0.41 (0.02)	0.52 (0.02)
	T	−0.72 (0.02)	−0.09 (0.01)	−0.01 (0.01)	—	0.41 (0.02)	0.51 (0.02)
	DPM	−1.12 (0.03)	−0.50 (0.02)	−0.41 (0.02)	−0.41 (0.02)	—	0.11 (0.01)
	MR	−1.23 (0.03)	−0.60 (0.02)	−0.52 (0.02)	−0.51 (0.02)	−0.11 (0.01)	—
Δ*Correct Decision Rate*	N	—	−0.04 (0.00)	−0.07 (0.00)	−0.07 (0.00)	−0.11 (0.01)	−0.13 (0.01)
	G-0	0.04 (0.00)	—	−0.03 (0.00)	−0.03 (0.00)	−0.07 (0.01)	−0.09 (0.01)
	G	0.07 (0.00)	0.03 (0.00)	—	0.00 (0.00)	−0.04 (0.00)	−0.06 (0.01)
	T	0.07 (0.00)	0.03 (0.00)	−0.00 (0.00)	—	−0.04 (0.00)	−0.06 (0.00)
	DPM	0.11 (0.01)	0.07 (0.01)	0.04 (0.00)	0.04 (0.00)	—	−0.02 (0.00)
	MR	0.13 (0.01)	0.09 (0.01)	0.06 (0.01)	0.06 (0.00)	0.02 (0.00)	—
Δ*Regret*	N	—	0.02 (0.00)	0.03 (0.00)	0.03 (0.00)	0.03 (0.00)	0.04 (0.00)
	G-0	−0.02 (0.00)	—	0.01 (0.00)	0.01 (0.00)	0.02 (0.00)	0.02 (0.00)
	G	−0.03 (0.00)	−0.01 (0.00)	—	−0.00 (0.00)	0.01 (0.00)	0.01 (0.00)
	T	−0.03 (0.00)	−0.01 (0.00)	0.00 (0.00)	—	0.01 (0.00)	0.01 (0.00)
	DPM	−0.03 (0.00)	−0.02 (0.00)	−0.01 (0.00)	−0.01 (0.00)	—	0.01 (0.00)
	MR	−0.04 (0.00)	−0.02 (0.00)	−0.01 (0.00)	−0.01 (0.00)	−0.01 (0.00)	—
Δ*Overestimation*	N	—	0.46 (0.02)	0.46 (0.02)	0.46 (0.02)	0.48 (0.02)	0.49 (0.02)
	G-0	−0.46 (0.02)	—	−0.00 (0.00)	−0.00 (0.00)	0.02 (0.00)	0.03 (0.00)
	G	−0.46 (0.02)	0.00 (0.00)	—	−0.00 (0.00)	0.02 (0.00)	0.03 (0.00)
	T	−0.46 (0.02)	0.00 (0.00)	0.00 (0.00)	—	0.02 (0.00)	0.03 (0.00)
	DPM	−0.48 (0.02)	−0.02 (0.00)	−0.02 (0.00)	−0.02 (0.00)	—	0.01 (0.00)
	MR	−0.49 (0.02)	−0.03 (0.00)	−0.03 (0.00)	−0.03 (0.00)	−0.01 (0.00)	—
Δ*MAE*	N	—	0.85 (0.02)	0.87 (0.02)	0.87 (0.02)	0.91 (0.02)	0.92 (0.02)
	G-0	−0.85 (0.02)	—	0.02 (0.00)	0.02 (0.00)	0.06 (0.00)	0.07 (0.00)
	G	−0.87 (0.02)	−0.02 (0.00)	—	0.00 (0.00)	0.04 (0.00)	0.06 (0.00)
	T	−0.87 (0.02)	−0.02 (0.00)	−0.00 (0.00)	—	0.04 (0.00)	0.06 (0.00)
	DPM	−0.91 (0.02)	−0.06 (0.00)	−0.04 (0.00)	−0.04 (0.00)	—	0.02 (0.00)
	MR	−0.92 (0.02)	−0.07 (0.00)	−0.06 (0.00)	−0.06 (0.00)	−0.02 (0.00)	—
Δ*Avg* 95% *CI Length*	N	—	3.97 (0.01)	4.01 (0.01)	3.97 (0.01)	4.39 (0.01)	4.46 (0.01)
	G-0	−3.97 (0.01)	—	0.05 (0.01)	0.00 (0.01)	0.43 (0.01)	0.50 (0.01)
	G	−4.01 (0.01)	−0.05 (0.01)	—	−0.04 (0.01)	0.38 (0.01)	0.45 (0.01)
	T	−3.97 (0.01)	−0.00 (0.01)	0.04 (0.01)	—	0.43 (0.01)	0.49 (0.01)
	DPM	−4.39 (0.01)	−0.43 (0.01)	−0.38 (0.01)	−0.43 (0.01)	—	0.07 (0.01)
	MR	−4.46 (0.01)	−0.50 (0.01)	−0.45 (0.01)	−0.49 (0.01)	−0.07 (0.01)	—

Notes. Numeric values denote comparisons between the row method and the column method (row – column for the metrics prefixed by Δ). Except for the *Win Rate* section, positive (negative) values represent comparisons in which the metric value for the row (column) exceeded that of the column (row). The values in parentheses denote the estimated standard error for these comparisons. The abbreviations in the column headings correspond to the naive, Gaussian-0, Gaussian, Student-*t*, Dirichlet process mixture, and metaregression methods, respectively. Avg, average; MAE, mean absolute error.

campaigns to website design changes, new product features, and updated personalization models. For each A/B test, we use the largest available sample (i.e., the most mature cohort). In most cases, these samples correspond with the data that were used to make implementation decisions.

6.1. Analysis Setup

Zillow receives revenue through various interrelated lines of business: new construction, rentals, sales leads for real estate agents, mortgages, products for sellers (e.g., promoted homes), and more. We might reasonably expect some lines of business to operate synergistically and others to result in cannibalization. To maintain confidentiality, we restrict attention to only two (undisclosed) lines of business, specifying $\mathbf{y}_t$ as estimates of the (rescaled) financial impact of the tested interventions on each individual line of business in terms of dollars per customer. If, for example, the two lines of business were new construction and rentals, $\mathbf{y}_t$

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

would consist of two components, $y_{t,nc}$ and $y_{t,r}$, representing the estimated per-customer effect of experiment t on revenue for new construction and rentals, respectively. Consequently, the natural definition for OEC_t in this case is $\text{OEC}_t = \mathbf{1}_2^\top \boldsymbol{\mu}_{\mathbf{y}t}$—the sum over the two lines of business as in Example 2.

The vector $\mathbf{x}_t$ consists of three short-term proxies for $\mathbf{y}_t$. Although we cannot disclose the specific definitions of these proxies, typical proxy metrics in this industry include measures of customer engagement, form-fill rates, and shorter-horizon measurements of $\mathbf{y}_t$. The primary purpose of including these proxies is to increase the model's posterior precision by including causal effect estimates that are predictive of OEC_t as quantified by the surrogacy measures given in Section 4.1.

We implemented all meta-analysis methods in Python using the collapsed Gibbs samplers described in Appendix B. For the priors, we set $\boldsymbol{\eta}_0 = \mathbf{0}_p$, $\boldsymbol{\Phi} = \mathbf{I}_p$, $\nu = 8$, and $\boldsymbol{\Psi} = 8\mathbf{I}_p$ for the Gaussian and T methods. For the T method, we assumed the prior $\log(\gamma) \sim \mathcal{N}(2,1)$ with a prior median of $e^2 \approx 7.4$ degrees of freedom. For the DPM, we set hyperpriors as follows: $\boldsymbol{\eta}_0 \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, $\boldsymbol{\Psi} \sim \mathcal{W}^{-1}(8\mathbf{I}_p, 8)$, $\phi \sim \mathcal{G}(1,1)$, $\alpha_0 \sim \mathcal{G}(4,4)$, and $\nu = 10$. Appendix C details a sensitivity analysis showing that the results in Section 6.2 are quite robust to specification of the prior distribution. We drew 30,000 samples for each method, with the first 10,000 samples discarded as warm-up iterations. The computation times were 1.35, 2.45, 4.73, and 142.53 minutes for the Gaussian-0, Gaussian, T, and DPM methods, respectively, on a cloud server with eight central processing units (CPUs) and 32 GB of memory. Appendix B.4 provides additional details on the computational scalability of our methods.

6.2. Sample-Split Predictive Analysis

To evaluate the methods' performance, we implement a sample-splitting approach, in which each A/B test is randomly divided at the customer level into two equal-sized subsets: a training set and a test set. We fit the methods on the training set, obtaining posterior-based estimates of OEC_t. We then compare these estimates with independent estimates generated from the test set. We generate these estimates using $\widehat{\text{OEC}}_t$, the naive estimator, because the naive method is asymptotically unbiased under only Assumptions 1 and 3; in particular, it does not require a correctly specified parametric model or even i.i.d. effects (Assumption 4). Consequently, it serves as an independent holdout benchmark for predictive evaluation in the empirical setting. In addition to point estimates, we produce 95% posterior predictive intervals for the naive test-set estimates of OEC_t for each method. These intervals allow us to assess whether the inferences from the meta-analysis methods are appropriately calibrated.

Table 3 summarizes the methods' predictive performance in terms of two performance metrics defined in Section 5.2: the *Win Rate* and *MLEQ*. The definitions of these metrics are modified slightly to accommodate the fact that we no longer have access to the ground-truth values of OEC_t; we replace these ground-truth values with the naive method's test-set estimates. We see that the four meta-analysis methods substantially outperform the naive approach in terms of the *Win Rate*, achieving lower prediction error in 69%–73% of A/B tests. The *MLEQ* metric produces the same conclusion as the *Win Rate*—that the meta-analysis methods substantially outperform the naive method. Among the meta-analysis methods, DPM and T perform best overall, with Gaussian and Gaussian-0 performing similarly but clearly better than the naive method. The DPM-naive comparison in the *MLEQ* block equals −0.29, meaning that the DPM's log absolute prediction error is 0.29 points lower than that of the naive method. Because this comparison is on the log absolute-error scale, exponentiating gives $\exp(-0.29) = 0.75$; equivalently, the DPM method reduces absolute prediction error by about 25% on the original scale. Quantifying the uncertainty in these performance metrics is challenging because the meta-analysis predictions are statistically dependent. The standard errors shown in

Table 3. Pair-Wise Method Comparisons for Predictive Performance

Metric	Method	Naive	Gaussian-0	Gaussian	T	DPM
Win Rate	Naive	—	0.27 (0.03)	0.27 (0.03)	0.28 (0.03)	0.31 (0.03)
	Gaussian-0	0.73 (0.03)	—	0.55 (0.04)	0.39 (0.04)	0.43 (0.04)
	Gaussian	0.73 (0.03)	0.45 (0.04)	—	0.40 (0.04)	0.43 (0.04)
	T	0.72 (0.03)	0.61 (0.04)	0.61 (0.04)	—	0.41 (0.04)
	DPM	0.69 (0.03)	0.57 (0.04)	0.57 (0.04)	0.59 (0.04)	—
Δ*Mean* log *Error Quotient*	Naive	—	0.22 (0.07)	0.22 (0.07)	0.28 (0.08)	0.29 (0.08)
	Gaussian-0	−0.21 (0.07)	—	0.00 (0.01)	0.06 (0.04)	0.08 (0.06)
	Gaussian	−0.22 (0.07)	0.00 (0.01)	—	0.06 (0.04)	0.08 (0.06)
	T	−0.27 (0.08)	−0.06 (0.04)	−0.06 (0.04)	—	0.02 (0.07)
	DPM	−0.29 (0.08)	−0.08 (0.06)	−0.08 (0.06)	−0.02 (0.07)	—

Notes. Pair-wise predictive performances of the five methods in the empirical application are shown with approximate standard errors in parentheses. The *Win Rate* section displays the proportion of tests in which the method listed in the row outperformed the method listed in the column. In the Δ*Mean* log *Error Quotient* section, a negative (positive) value indicates that the row (column) method outperformed the column (row) method.

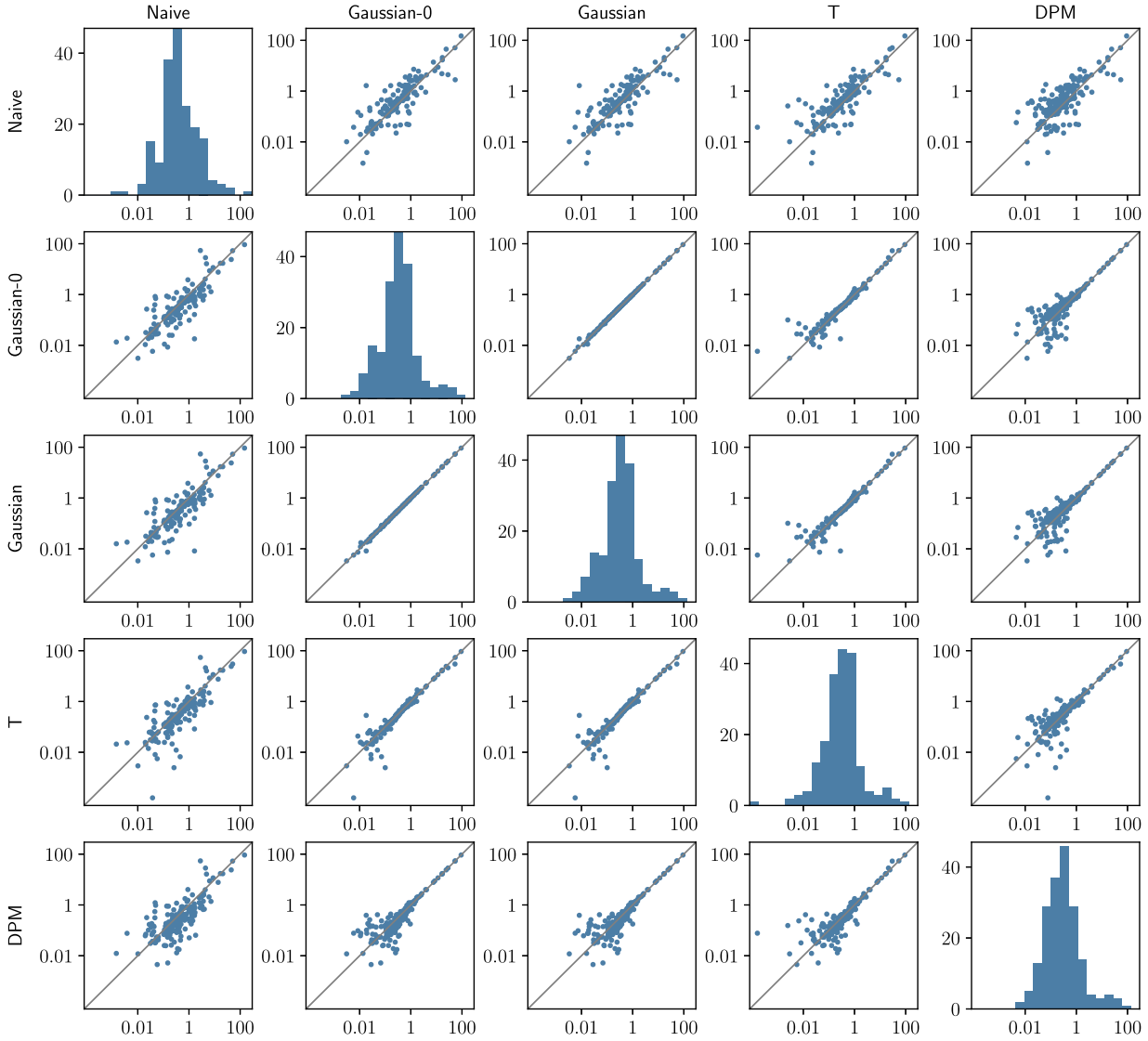
Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

parentheses in Table 3 are calculated assuming independence of the predictions, so they should be interpreted as rough approximations. We address the uncertainty quantification more rigorously in our simulation study in Section 5 using independent replications. Compared with the results in Table 2, we see that the predictive nature of this comparison dilutes the differences between methods because (i) the "true value" in this analysis is a noisy estimate of OEC_t and (ii) the estimates are formed using only 50% of the data.

Figure 4 visually compares the methods' absolute prediction errors in a pair-wise fashion. Below the diagonal in Figure 4, methods listed in each row can visually be seen to outperform the method listed in the column when the points fall below the 45° line. These plots in Figure 4 provide visual confirmation of the relative performances shown in Table 3. In particular, we see that the meta-analysis methods typically produce smaller absolute prediction errors than the naive method. The large differences in scale can be observed by noting that the absolute prediction errors shown are plotted on the log scale, with most values falling between 0.01 and 100.

Figure 5 plots the ranks of the predictions against the ranks of the prediction errors. Ideally, we would see no linear relationship between these ranks, indicating that the inferences are unbiased in a conditional sense (i.e., given a subset of tests having estimated OEC_t equal to

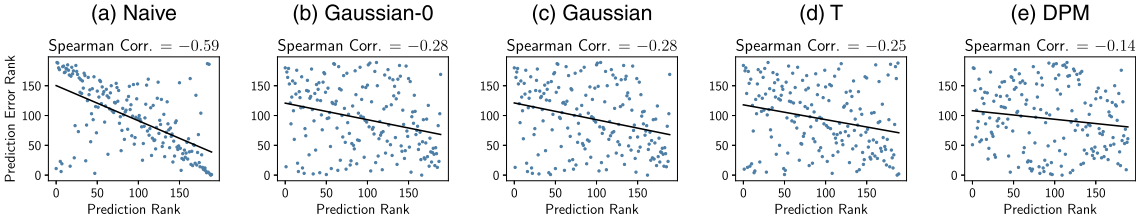
Figure 4. (Color online) Absolute Prediction Error by Method



Notes. Each coordinate represents the absolute prediction error for the corresponding method in the empirical application. The diagonal lines represent exactly equal performance. The DPM's prediction errors are the smallest, indicating that it produced the most accurate estimates of OEC_t on average.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Figure 5. (Color online) Scatterplots Depicting the Association Between Ranks of the Predictions and Their Errors



Notes. The large negative trend in panel (a) indicates regression to the mean. In contrast, the flatter regression lines and smaller absolute values of the Spearman correlation (Corr.) coefficient in the remaining panels indicate that the meta-analysis methods mitigate this phenomenon. The DPM method in panel (e) in particular shows the least evidence of regression to the mean among the fitted models. (a) Naive. (b) Gaussian-0. (c) Gaussian. (d) T. (e) DPM.

$u \in \mathbb{R}$, the conditional expectation of OEC_t for these tests is, in fact, equal to the estimate, u). Although this seems to be a natural methodological requirement, the naive approach does not satisfy it because of regression to the mean. This shortcoming is visually apparent in panel (a) of Figure 5 with the strong negative correlation between the predictions and errors ($\rho = -0.59$). In common practice, A/B tests with large, statistically significant OEC_t estimates are implemented and internally reported. The unfortunate implication is that the impact of these "winning" tests is typically overstated. Bayesian methods provide a natural fix for this problem by performing conditional inference; under correct model specification, the posterior mean possesses a conditional unbiasedness property by design. This property is evident in panels (b)–(e) of Figure 5, which show much lower correlation between prediction ranks and error ranks compared with the naive approach ($|\rho| < 0.3$ in all four cases). The DPM model achieves the greatest reduction as evidenced by the relatively flat line in panel (e) of Figure 5 and a small Spearman rank correlation of -0.14. This finding suggests that the DPM method is likely modeling the true prior distribution, $P_{\boldsymbol{\theta}}$, more accurately than the other methods. It effectively debiases the naive estimates ($\mathbf{z}_t$) by accurately modeling the heterogeneity in the true effects ($\boldsymbol{\mu}_t$) via its flexible representation of the prior distribution as a mixture of inferred clusters. It regularizes each estimate just enough that its expected error is near zero regardless of the value of the naive estimate. This property can also be seen in the overestimation metric in Section 5.2, which is almost exactly zero for the DPM method.

Table 4 compares the performance of the methods' prediction intervals. All methods produce near-nominal coverage. However, the meta-analysis methods produce shorter intervals. This is especially true of the DPM method, which produces median interval lengths that are about 25% shorter than those of the naive method, indicating substantial benefit from including a prior distribution for the $\boldsymbol{\mu}_t$ values. Again, the predictive nature of this comparison dilutes the difference in interval width.

Table 4. Performance of 95% Prediction Intervals

	N	G-0	G	T	DPM
Coverage	0.95	0.94	0.95	0.96	0.96
Median length	2.06	1.78	1.78	1.83	1.55
Mean length	10.21	7.38	7.38	7.75	7.21

Notes. The 95% prediction interval coverage is shown for the methods tested in the empirical application: naive, Gaussian-0, Gaussian, Student-t, and Dirichlet process mixture. All methods achieve approximately nominal coverage. However, the meta-analysis methods produce much shorter prediction intervals on average.

6.3. Full-Data Analysis

In addition to the sample-split predictive analysis, we also applied the DPM method to the full data set. To evaluate Assumption 4, we used the DPM model to compute posterior predictive p-values, checking for correlation between the entries of $\boldsymbol{\mu}_t$ (and their absolute values) and the corresponding variance entries in $\boldsymbol{\Sigma}_t$. The smallest posterior predictive p-value was 0.05. With 10 comparisons, this represents only minimal evidence against Assumption 4. The rest of the analysis provides insights into the inner workings of the DPM model and the latent structure of the causal effects across tests.

Table 5 displays estimates of the surrogacy measures introduced in Section 4.1. Because the DPM model produces test-specific estimates of these measures, we

Table 5. Aggregate Surrogacy Measures

	μ_1	μ_2	$\text{OEC}_t := \mu_1 + \mu_2$
R^2	0.52 (0.18–0.90)	0.78 (0.42–0.98)	0.46 (0.13–0.81)
TVR	0.16 (0.08–0.24)	0.10 (0.04–0.19)	0.15 (0.08–0.24)
PVR	0.18 (0.10–0.27)	0.13 (0.06–0.22)	0.17 (0.10–0.26)
SVR	0.91 (0.82–0.97)	0.88 (0.79–0.95)	0.93 (0.84–0.97)

Notes. Posterior means for aggregated surrogacy measures from the DPM model (95% credible intervals) are shown. The surrogacy measures (defined in Sections 4.1 and 4.3) are averaged over all 190 tests. Notably, the low values of PVR relative to SVR indicate that in this application, most of the variance reduction arises from prior regularization of the long-run outcomes themselves, with the short-term proxies providing additional but comparatively smaller gains.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

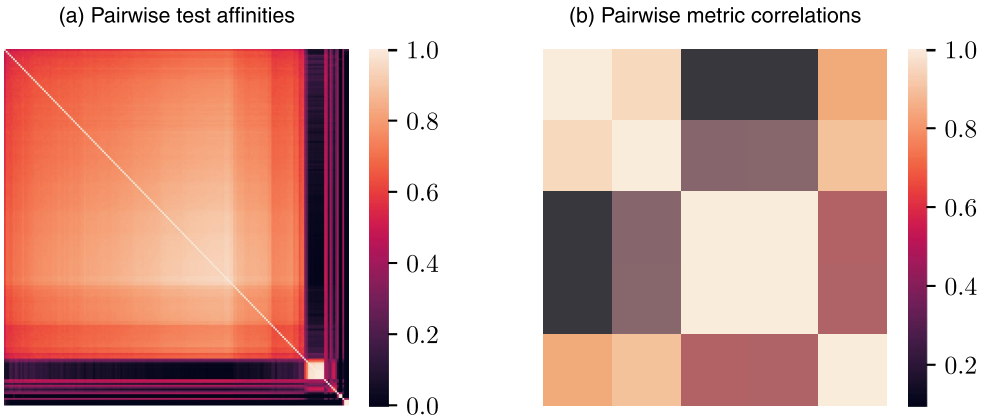
average their values across tests to obtain estimates of the overall strength of surrogacy. The estimated values of R^2 fall in the range from 0.46 to 0.78, indicating moderately strong surrogacy. Effects on the surrogates ($\mathbf{x}_t$) do, in fact, predict effects on the KPIs ($\mathbf{y}_t$); however, they do so imperfectly. The R^2 value for OEC_t is smaller than that of both μ_1 and μ_2, indicating that μ_1 and μ_2 show some degree of internal surrogacy. The estimates of TVR are in the range from 0.10 to 0.16, indicating a reduction in the posterior standard deviation of over 60% compared with the naive approach. Achieving a comparable reduction using the naive method alone would require increasing the sample size by a factor of 1/TVR; here, this corresponds with a factor of 6–10, which is not achievable because these tests are continuously run in sequence across different teams. These metrics can also be used to expand and refine the list of included surrogates. For instance, Table 5 indicates that the surrogates for μ_2 have larger R^2 values than those for μ_1, suggesting that the analysis may benefit relatively more from adding a strong surrogate for μ_1 compared with adding a strong surrogate for μ_2.

Figure 6 displays heat maps that visualize the inner workings of the DPM model. The posterior mean for the number of clusters is 7.2 with a 95% credible interval of 4–12. Panel (a) of Figure 6 plots posterior probabilities that pairs of tests are clustered together. These probabilities represent the model's posterior belief that the tests arise from the same distribution. We used hierarchical clustering methods (Murtagh and Contreras 2012) to group and sort the tests.[10] Doing so reveals the existence of several small, distinct clusters that are internally homogeneous but distinct from the other clusters. We applied the DPM method to an enlarged set of business metrics and obtained similar results, although with more distinct clusters. Upon inspecting these clusters, we found many cases where the DPM method had clustered tests involving a similar intervention or platform. For instance, one cluster of tests primarily affected the search experience for rental properties, and another involved only tests implemented on mobile devices. In practice, these model-based clusterings could be used to identify experiment-level metadata that could be used within a metaregression model.

The values of $\boldsymbol{\eta}_k$ and $\boldsymbol{\Omega}_k$ provide valuable strategic insights, such as the average effect of these interventions and how these effects correlate across metrics. An example of the latter is that the surrogacy relationships may differ across these clusters relative to the main cluster. The fact that the DPM model has access only to the values of $\mathbf{z}_t$ and $\boldsymbol{\Sigma}_t$ indicates that these groupings are apparent from the test results themselves. In addition to improving our estimates of $\boldsymbol{\mu}_t$, these latent groupings could be used to learn the "profile" of certain types of tests, which could be used to ideate and prioritize future interventions. In analyzing the tth test, the latent groupings could also inform decision making by comparing results across similar tests.

Panel (b) of Figure 6 displays estimated correlations among the components of $\boldsymbol{\mu}_t$. To further preserve confidentiality, the order of the metrics has been permuted. The plot in panel (b) of Figure 6 reveals that two pairs of metrics show strong positive correlation within pairs but are nearly uncorrelated between pairs. Additionally, the last metric shown is positively correlated with all four of the other metrics. The fact that all five metrics are, to varying degrees, positively correlated suggests that these metrics tend to have synergistic effects—improvements in one metric are associated with improvements in the others. Thus, in addition to offering improved effect estimates, the meta-analysis methods also uncover interpretable insights regarding the distribution of causal effects. As an example, suppose we learn that the treatment effects on a form-fill metric

Figure 6. (Color online) Heat Maps Showing Estimated Latent Structure in the 190 A/B Tests



Notes. Panel (a) displays posterior probabilities that tests are clustered together (pair-wise affinities), revealing the existence of several small, distinct clusters. Panel (b) displays pair-wise estimated Pearson correlations for the latent effects ($\boldsymbol{\mu}_t$) across experiments. Each location on the axis in panel (b) corresponds with one of the elements of $\boldsymbol{\mu}_t$; the order of the metrics has been permuted to preserve confidentiality. (a) Pair-wise affinities. (b) Pair-wise metric correlations.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

and a conversion metric are highly correlated across experiments. This finding implies that interventions improving the form-fill rate also tend to improve downstream conversions, which may influence the firm's allocation of experimentation resources.

In the biostatistics literature on clinical trials, correlations like these are sometimes used to justify using one surrogate metric in place of another that is noisier or more difficult to measure. Although this strategy could be applied in A/B testing as well, we generally recommend selecting a target estimand (OEC_t) that closely aligns with long-term business goals and applying our meta-analysis methods to frame all subsequent inferences in terms of this estimand; the resulting strategy is analogous to the procedures proposed in Elliott et al. (2015) to avoid the so-called "surrogacy paradox," in which a strong surrogate produces incorrect inferences for a particular intervention. A key benefit of the experimental meta-analysis approach over purely observational (e.g., data-mining) methods is that the effects are causal by design—not a misleading artifact of noncausal associations. Thus, in addition to improving A/B test decision making, the meta-analysis framework can provide crucial business insights that may be missed or badly biased in an observational analysis.

7. Discussion

This section provides a brief summary of our proposed framework, the relevant managerial considerations, and possible extensions.

7.1. Methodological and Results Summary

The primary contribution of this work is a meta-analytic framework for estimating the long-term value of A/B tests that (i) replaces additional causal surrogacy assumptions with an assumption that comparable experiments have exchangeable causal-effect vectors, (ii) leverages cross-metric relationships for increased efficiency, and (iii) is computationally scalable because of its reliance on summary statistics. Additionally, we demonstrate via simulation and an empirical application to 190 A/B tests in real estate technology that the proposed framework provides a large improvement over single-test analyses. Although our application is in real estate technology specifically, the framework is generally applicable to settings with many experiments that (i) share metric definitions, (ii) present little or no cross-experiment interference, and (iii) arise from a common domain so that the shared-effect distribution assumption in Assumption 4 or its conditional version in Assumption 5 is credible. Conversely, the framework is less suitable when experiments interfere significantly or differ sharply in domain, customer population, or metric definition unless these differences can be modeled explicitly through metaregression or careful filtering of the experiment archive.

We have presented three basic meta-analytic models: Gaussian, T, and DPM. We recommend informing the choice of the meta-analytic model via modeling diagnostics, including posterior model checking and information criteria (Spiegelhalter et al. 2002, Watanabe 2010). Because of its flexibility, we generally expect the DPM to outperform the Gaussian and T models; however, the latter two models can be fit much more quickly. In practice, analysts also have the flexibility to generalize well beyond these models, specifying metaregression models that allow the effect distribution to vary based on experiment attributes. When relevant sources of heterogeneity are omitted (e.g., differences in customer segments, product types, or experimental contexts), the model may be well calibrated only on average across experiments and can perform poorly for specific subsets of tests. In general, we expect that more heterogeneous experiments will require more flexible models to satisfy Assumption 4 or Assumption 5. Appendix A discusses additional practical considerations for applying our framework.

7.2. Managerial Implications

This work carries two primary managerial implications. First, Sections 5 and 6 uncovered a large statistical benefit in adopting meta-analytic methods. Our proposed framework offers a satisfying solution to the dilemma of choosing which business metrics or portion of the business funnel (top versus middle versus bottom) to prioritize. Our framework enables managers to define success in terms of bottom-of-funnel results (purchases, customer retention, etc.) and simultaneously reap the statistical benefits of top-of-funnel proxies (click-through rates, form-fill rates, etc.). These proxy metrics are relevant to decision making to the extent that causal effects on the proxies correlate with causal effects on strategic bottom-of-funnel outcomes. As a result, our meta-analysis methods provide a data-driven alternative to the industry practice of "OEC by committee," in which business leaders manually select a set of metrics to prioritize. The meta-analysis methods also provide an elegant solution to the challenges of simpler OEC methods, including (i) the need to customize the OEC to different lines of business or software platforms, (ii) the inability to measure certain metrics for some tests because of technical constraints, and (iii) the sensitivity of the OEC definition to the time frame over which metrics are measured.

Second, our framework enables rigorous tracking of cumulative long-term experimental impacts, even when tests are selected for implementation or reporting based on large positive estimated effects. Because of the statistical phenomenon of regression to the mean, this common decision-making practice can substantially overstate

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

cumulative impacts; see panel (a) of Figure 5 and our results on the overestimation metric in Section 5. Our meta-analytic framework addresses this problem by "shrinking" estimates toward regions of high prior probability, substantially reducing regression to the mean effects and the overstatement induced by selecting tests with large positive naive estimates $\mathbf{z}_t$. In practice, this feature of our framework can facilitate decision making surrounding organizational resource allocation and strategic priorities by producing reliable estimates of the financial value of implemented A/B tests.

7.3. Extensions

We present three common meta-analysis methods ranging from simple parametric models to flexible nonparametric alternatives. However, the statistical literature on meta-analysis and random-effects models is vast and includes many alternatives that could be applied within our OEC framework, including parametric and nonparametric empirical Bayes methods (Higgins et al. 2009, Efron 2014, Koenker and Mizera 2014). The Bayesian solutions presented here are notable for their extensibility and ability to provide actionable inferences, including surrogacy measures, explicit estimates of $P_{\boldsymbol{\theta}}$, and estimated clusterings.

Another promising direction is to develop methods that extrapolate effects beyond a short-term treatment window. Our framework can estimate the long-term impact of *short-term treatments*, averaging over any trends or business decisions between the test end date and the end of the measurement period. However, in some business settings, the true goal may be to estimate the long-term impact of *long-term treatments*, such as the effect on 12-month revenue of adding (versus not adding) a new feature over a 6-month period. This extension requires a different set of methods that effectively forecast the long-term effects using short-term proxies (Huang et al. 2026). Future work could consider how to combine these methods with the meta-analysis perspective presented here.

Acknowledgments

The authors thank Eric Ringger, Ryan Williamson, and other members of the data science and artificial intelligence organizations at Zillow for their important contributions to this research. The authors also thank the anonymous reviewers, the associate editor, and the senior editor for providing invaluable feedback that greatly improved the manuscript. Stephanie A. Nail, Terran Melconian, and Xan Vongsathorn are full-time employees at Zillow. Easton Huch was employed as an applied scientist intern while contributing to this study. Zillow has complete rights to review and require changes to this paper for any material objectionable to its interests, including the removal of intellectual property and trade secrets.

Appendix A. Additional Practical Considerations

In practice, tests with more than two arms are common. To accommodate these tests, we recommend stacking the effect estimates for the noncontrol arms into a single vector: $\mathbf{z}_t := (\mathbf{z}_{t1}^\top \dots \mathbf{z}_{ta_t}^\top)^\top$, where $a_t \in \mathbb{N}$ is the total number of noncontrol arms and $\mathbf{z}_{tj}$ is the vector of effect estimates for the jth such arm. Similarly, the joint covariance matrix of these effect estimates can be formed as

$$\boldsymbol{\Sigma}_t := \begin{bmatrix} \boldsymbol{\Sigma}_{t11} & \dots & \boldsymbol{\Sigma}_{t1a_t} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{ta_t1} & \dots & \boldsymbol{\Sigma}_{ta_ta_t} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{tij} := \mathrm{Cov}(\mathbf{z}_{ti}, \mathbf{z}_{tj})$. Under the assumption $\boldsymbol{\mu}_{tj} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Omega})$ for $j \in [a_t]$, the BLUP for the tth test generalizes to

$$(\mathbf{I}_{a_t} \otimes \boldsymbol{\Omega}^{-1} + \boldsymbol{\Sigma}_t^{-1})^{-1} \{\mathbf{1}_{a_t} \otimes (\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}) + \boldsymbol{\Sigma}_t^{-1}\mathbf{z}_t\}.$$

The remaining Gibbs updates can be generalized in much the same fashion.

Another common practical challenge is that certain metrics may be missing because of (i) an incomplete measurement period or (ii) the inability to track certain outcomes in specific types of tests. In these cases, the Gibbs samplers can be modified as described in (6) and Appendix B. A simple (albeit less elegant) alternative is to fill the missing submatrix of $\boldsymbol{\Sigma}_t$ with a large diagonal matrix and set the missing elements of $\mathbf{z}_t$ to zero.

Operationally, we recommend maintaining a centralized archive of experiment-level summaries $(\mathbf{z}_t, \boldsymbol{\Sigma}_t)$ with standardized metric definitions and time windows, running periodic model checks, and adopting more flexible methods (nonparametric models and/or metaregression) when model assumptions are not satisfied. Teams can refit these models regularly (e.g., daily) using the most up-to-date estimates available, extracting posterior summaries of OEC_t for in-progress tests. This process can be complemented by comparing early estimates with "mature estimates" available after all metrics are observed.

Appendix B. Gibbs Sampling Algorithms

This section provides the Gibbs samplers for the Gaussian, Student-t, and DPM models.

B.1. Gaussian Gibbs Sampler

For simplicity, we assume independent Gaussian and inverse-Wishart priors for $\boldsymbol{\eta}$ and $\boldsymbol{\Omega}$, respectively. Denoting the prior scale matrix and degrees of freedom as $\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}$ and $\nu > p-1$, respectively, the complete conditional for $\boldsymbol{\Omega}$ is $\mathcal{W}^{-1}(\boldsymbol{\Psi}_T, \nu_T)$, where $\mathcal{W}^{-1}$ denotes an inverse-Wishart distribution and the posterior parameters are

$$\boldsymbol{\Psi}_T = \boldsymbol{\Psi} + \sum_{t\in[T]} (\boldsymbol{\mu}_t - \boldsymbol{\eta})(\boldsymbol{\mu}_t - \boldsymbol{\eta})^\top, \quad \nu_T = \nu + T. \tag{B.1}$$

We denote the prior mean and prior variance of $\boldsymbol{\eta}$ as $\boldsymbol{\eta}_0 \in \mathbb{R}^p$ and $\boldsymbol{\Phi} \in \mathbb{R}^{p\times p}$, respectively. The parameters $\boldsymbol{\eta}$ and $\{\boldsymbol{\mu}_t\}_{t\in[T]}$ can be blocked together in the Gibbs sampler by first integrating out the $\{\boldsymbol{\mu}_t\}_{t\in[T]}$ to improve mixing, resulting in a "collapsed" Gibbs sampler (Liu et al. 1994). The marginalized complete conditional distribution for $\boldsymbol{\eta}$ is $\mathcal{N}(\boldsymbol{\eta}_T, \boldsymbol{\Phi}_T)$ with posterior parameters

$$\boldsymbol{\Phi}_T = \left[\boldsymbol{\Phi}^{-1} + \sum_{t\in[T]} (\boldsymbol{\Omega} + \boldsymbol{\Sigma}_t)^{-1}\right]^{-1}, \tag{B.2a}$$

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

$$\boldsymbol{\eta}_T = \boldsymbol{\Phi}_T\left[\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}_0 + \sum_{t\in[T]}(\boldsymbol{\Omega}+\boldsymbol{\Sigma}_t)^{-1}\mathbf{z}_t\right]. \tag{B.2b}$$

Similar to (6), we can handle partial observations by replacing $(\boldsymbol{\Omega}+\boldsymbol{\Sigma}_t)^{-1}$ with $\mathbf{A}^\top[\mathbf{A}(\boldsymbol{\Omega}+\boldsymbol{\Sigma}_t)\mathbf{A}^\top]^{-1}\mathbf{A}$. After sampling $\boldsymbol{\eta}$, $\{\boldsymbol{\mu}_t\}_{t\in[T]}$ can then be sampled according to (5). Algorithm B.1 summarizes the Gibbs sampler.

In specifying the hyperprior, we recommend setting $\boldsymbol{\eta}_0 = \mathbf{0}_p$, corresponding to a prior belief that the average effect is zero. Because $\boldsymbol{\eta}$ is well identified from the data, we recommend setting $\boldsymbol{\Phi}$ to a diagonal matrix with large nonzero entries. The prior for $\boldsymbol{\Omega}$ can likewise be set in a weakly informative fashion, selecting a small value of ν (such as p) and a corresponding value of $\boldsymbol{\Psi}$ that results in a prior mode that roughly matches a priori beliefs regarding the magnitude of the true effects ($\boldsymbol{\mu}_t$).

Algorithm B.1 (Gaussian Gibbs Scan)
Input: $\{\boldsymbol{\mu}_t, \mathbf{z}_t, \boldsymbol{\Sigma}_t\}_{t\in[T]}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\eta}_0, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \nu$
Sample $\boldsymbol{\Omega}$ from $\mathcal{W}^{-1}(\boldsymbol{\Psi}_T, \nu_T)$ with $\boldsymbol{\Psi}_T, \nu_T$ defined in (B.1).
Sample $\boldsymbol{\eta}$ from $\mathcal{N}(\boldsymbol{\eta}_T, \boldsymbol{\Phi}_T)$ with $\boldsymbol{\eta}_T, \boldsymbol{\Phi}_T$ defined in (B.2).
for $t \in [T]$ **do**
  Sample $\boldsymbol{\mu}_t$ from $\mathcal{N}(\boldsymbol{\eta}_t, \boldsymbol{\Omega}_t)$ with $\boldsymbol{\eta}_t, \boldsymbol{\Omega}_t$ defined in (5b).
end for

B.2. Student-*t* Gibbs Sampler

Unlike the Gaussian model, the multivariate Student-*t* model is not conjugate given $\boldsymbol{\theta}$. However, it is conditionally conjugate if we introduce latent variables $\{d_t\}_{t\in[T]}$ as follows:

$$d_t|\gamma \overset{iid}{\sim} \chi^2(\gamma), \quad \boldsymbol{\mu}_t|d_t, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{N}\left(\boldsymbol{\eta}, \frac{\gamma}{d_t}\boldsymbol{\Omega}\right).$$

This approach is a form of Bayesian data augmentation (Tanner and Wong 1987, van Dyk and Meng 2001). It maintains the assumption of (8) and introduces simple conditionally conjugate updates for d_t and $\boldsymbol{\mu}_t$. The Gibbs sampler of Appendix B.1 can be extended to fit this model by introducing a sampling block for γ and $\{d_t\}_{t\in[T]}$. We recommend first integrating out $\{d_t\}_{t\in[T]}$ and using a Metropolis–Hastings step to sample γ. Conditional on γ, the latent χ^2 variables $\{d_t\}_{t\in[T]}$ can then be sampled as $d_t|\cdot \overset{ind}{\sim} \mathcal{G}\{(\gamma+p)/2, \beta/2\}$, where $\mathcal{G}$ denotes a gamma distribution and the rate parameter is $\beta = 1 + (\boldsymbol{\mu}_t - \boldsymbol{\eta})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\eta})/\gamma$. The remaining Gibbs blocks are similar to those in Appendix B.1 with the following rescalings by the ratio γ/d_t:

$$\boldsymbol{\Psi}_T = \boldsymbol{\Psi} + \sum_{t\in[T]}\frac{d_t}{\gamma}(\boldsymbol{\mu}_t - \boldsymbol{\eta})(\boldsymbol{\mu}_t - \boldsymbol{\eta})^\top,$$

$$\boldsymbol{\Phi}_T = \left[\boldsymbol{\Phi}^{-1} + \sum_{t\in[T]}\left(\frac{\gamma}{d_t}\boldsymbol{\Omega} + \boldsymbol{\Sigma}_t\right)^{-1}\right]^{-1},$$

$$\boldsymbol{\eta}_T = \boldsymbol{\Phi}_T\left[\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}_0 + \sum_{t\in[T]}\left(\frac{\gamma}{d_t}\boldsymbol{\Omega} + \boldsymbol{\Sigma}_t\right)^{-1}\mathbf{z}_t\right].$$

The posterior parameters $\boldsymbol{\eta}_t$ and $\boldsymbol{\Omega}_t$ are adapted in a similar fashion, replacing $\boldsymbol{\Omega}$ with $\gamma\boldsymbol{\Omega}/d_t$. As with the multivariate Gaussian model, the Gibbs steps can easily be generalized to the setting in which only $\mathbf{A}\mathbf{z}_t$ is observed.

We recommend setting the hyperparameters $\boldsymbol{\eta}_0$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and ν as described in Appendix B.1. We recommend specifying a prior for γ such that it includes values ranging from high kurtosis ($\gamma \approx 5$) to nearly Gaussian ($\gamma \approx 10$) with high prior probability.

B.3. DPM Gibbs Sampler

Gibbs sampling is a particularly common computational approach for DPM models because their infinite parameter space and multimodality make them less amenable to HMC. We fit our DPM model via a collapsed Gibbs sampler (Neal 2000, algorithm 3). The sampler iteratively updates the cluster assignments, c_t, based on the posterior predictive density of each cluster, $g_k(\boldsymbol{\mu}; \{\boldsymbol{\mu}_t\}_{t\in\mathbb{S}_k})$, available in closed form as the following multivariate Student-*t* density:

$$\tau\left[\boldsymbol{\mu}; \boldsymbol{\eta}_k, \left(1+\frac{1}{\phi_k}\right)\frac{\boldsymbol{\Psi}_k}{\nu_k - p + 1}, \nu_k - p + 1\right], \tag{B.3}$$

where $\mathbb{S}_k := \{t \in [T] : c_t = k\}$ is the set of indices for members of cluster k. Denoting the number of cluster members as $T_k := \#\mathbb{S}_k$, the posterior parameters in (B.3) are

$$\nu_k := \nu + T_k, \quad \phi_k := \phi + T_k, \quad \boldsymbol{\eta}_k := \frac{\phi\boldsymbol{\eta}_0 + T_k\bar{\boldsymbol{\mu}}_k}{\phi + T_k}, \tag{B.4a}$$

$$\boldsymbol{\Psi}_k := \boldsymbol{\Psi} + \mathbf{S}_k + \frac{\phi T_k}{\phi + T_k}(\bar{\boldsymbol{\mu}}_k - \boldsymbol{\eta}_0)(\bar{\boldsymbol{\mu}}_k - \boldsymbol{\eta}_0)^\top, \tag{B.4b}$$

$$\bar{\boldsymbol{\mu}}_k := \frac{1}{T_k}\sum_{t\in\mathbb{S}_k}\boldsymbol{\mu}_t, \quad \mathbf{S}_k := \sum_{t\in\mathbb{S}_k}(\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_k)(\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_k)^\top. \tag{B.4c}$$

The Gibbs sampler reassigns points to existing clusters with probabilities proportional to $T_k g_k(\boldsymbol{\mu}_t; \{\boldsymbol{\mu}_u\}_{u\in\mathbb{S}_k\setminus\{t\}})$ and to a new cluster with probability proportional to $\alpha_0 g_{K+1}(\boldsymbol{\mu}_t; \varnothing)$. This loop is the most computationally costly portion of the Gibbs sampler, but it can be optimized by updating the parameters in (B.4) via efficient rank-1 updates. After updating all c_t in this manner, we sample $\boldsymbol{\Omega}_k$ from $\mathcal{W}^{-1}(\boldsymbol{\Psi}_k, \nu_k)$ using the full set $\mathbb{S}_k$. We then update $\boldsymbol{\eta}_k$ with a collapsed Gibbs step with the parameters modified from (B.2) as follows:

$$\boldsymbol{\Phi}_{Tk} = \left[(\phi\boldsymbol{\Omega}_k)^{-1} + \sum_{t\in\mathbb{S}_k}(\boldsymbol{\Omega}_k + \boldsymbol{\Sigma}_t)^{-1}\right]^{-1}, \tag{B.5a}$$

$$\boldsymbol{\eta}_{Tk} = \boldsymbol{\Phi}_{Tk}\left[(\phi\boldsymbol{\Omega}_k)^{-1}\boldsymbol{\eta}_0 + \sum_{t\in\mathbb{S}_k}(\boldsymbol{\Omega}_k + \boldsymbol{\Sigma}_t)^{-1}\mathbf{z}_t\right]. \tag{B.5b}$$

Subsequently, we update $\boldsymbol{\mu}_t$ according to (5), replacing $\boldsymbol{\Omega}$ and $\boldsymbol{\eta}$ with the cluster-specific parameters $\boldsymbol{\Omega}_{c_t}$ and $\boldsymbol{\eta}_{c_t}$, respectively. As with the multivariate Gaussian and Student-*t* methods, these steps can easily be adapted to the setting in which only $\mathbf{A}\mathbf{z}_t$ is observed via slight modifications to the updates for $\{\boldsymbol{\Omega}_k\}_{k\in[K]}$, $\{\boldsymbol{\eta}_k\}_{k\in[K]}$, and $\{\boldsymbol{\mu}_t\}_{t\in[T]}$. The GEM parameter, α_0, can be updated via Metropolis–Hastings or the conditionally conjugate data augmentation strategy given in Escobar and West (1995). The prior parameters $\boldsymbol{\Psi}$, $\boldsymbol{\eta}_0$, ϕ, and ν can also be updated using conditionally conjugate priors or a Metropolis–Hastings step. Algorithm B.2 summarizes the full Gibbs scan.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Algorithm B.2 (DPM Gibbs Scan)

Input: $\{c_t, \boldsymbol{\mu}_t, \mathbf{z}_t, \boldsymbol{\Sigma}_t\}_{t\in[T]}, \{\boldsymbol{\eta}_k, \boldsymbol{\Omega}_k\}_{k=1}^{K}, \alpha_0, \nu, \phi, \boldsymbol{\Psi}, \boldsymbol{\eta}_0$

$K \leftarrow \max_{t\in[T]} c_t$

$\mathbb{S}_k \leftarrow \{t \in [T] : c_t = k\}$ for all $k \in [K+1]$

for $t \in [T]$ **do**

Remove t from $\mathbb{S}_{c_t}$

if $\mathbb{S}_{c_t} = \varnothing$ **then**

$K \leftarrow K - 1$

$c_u \leftarrow c_u - 1$ for all $u \in [T] \setminus \{t\}$ with $c_u > c_t$

end if

Sample a new value for c_t using g as defined in (B.3) with $k \in [K]$:

$$\Pr(c_t = k) \propto T_k g_k(\boldsymbol{\mu}_t; \{\boldsymbol{\mu}_u\}_{u\in\mathbb{S}_k}),$$
$$\Pr(c_t = K+1) \propto \alpha_0 g_{K+1}(\boldsymbol{\mu}_t; \varnothing)$$

$K \leftarrow \max(K, c_t)$ and append t to $\mathbb{S}_{c_t}$

end for

Sample $\{\boldsymbol{\Omega}_k\}_{k\in[K]}$ from $\mathcal{W}^{-1}(\boldsymbol{\Psi}_k, \nu_k)$ with $\boldsymbol{\Psi}_k, \nu_k$ defined in (B.4)

Sample $\{\boldsymbol{\eta}_k\}_{k\in[K]}$ from $\mathcal{N}(\boldsymbol{\eta}_{Tk}, \boldsymbol{\Phi}_{Tk})$ with $\boldsymbol{\eta}_{Tk}, \boldsymbol{\Phi}_{Tk}$ defined in (B.5)

Sample $\{\boldsymbol{\mu}_t\}_{t\in[T]}$ according to (5), substituting $\boldsymbol{\Omega}_{c_t}$ for $\boldsymbol{\Omega}$ and $\boldsymbol{\eta}_{c_t}$ for $\boldsymbol{\eta}$

(Optional) Sample α_0 via Metropolis–Hastings or the data augmentation method given in Escobar and West (1995)

(Optional) Sample $\nu, \phi, \boldsymbol{\Psi}, \boldsymbol{\eta}_0$ via conditionally conjugate distributions or Metropolis–Hastings

We recommend setting $\alpha_0 \in (0.5, 2)$ or estimating it with a prior centered in the same interval. Whenever feasible, we recommend specifying priors for ν, ϕ, $\boldsymbol{\Psi}$, and $\boldsymbol{\eta}_0$. The prior for ν and $\boldsymbol{\Psi}$ can be set as explained in Appendix B.1. We recommend setting the prior for ϕ to a dispersed distribution on $\mathbb{R}^+$ that includes the value of one, such as $\mathcal{G}(1,1)$. We recommend setting the prior for $\boldsymbol{\eta}_0$ to a dispersed multivariate Gaussian distribution centered at $\mathbf{0}_p$, similar to the above recommendation for setting $\boldsymbol{\eta}_0$ and $\boldsymbol{\Phi}$ in the Gaussian and T methods.

B.4. Computational Scalability

With the cost of cloud CPU hours below \$0.05, the computation time required for our methods represents a trivial cost for most technology firms. We generally expect the cost of computing $\mathbf{z}_t$ and $\boldsymbol{\Sigma}_t$ to be much higher than the cost of fitting these models. However, standard univariate A/B testing procedures already require computation of $\mathbf{z}_t$ and $\mathrm{diag}(\boldsymbol{\Sigma}_t)$, so the only additional required computations are the covariance terms (the off-diagonal elements) of $\boldsymbol{\Sigma}_t$.

The samplers scale approximately linearly in T, so the methods could easily scale to thousands of tests, especially the Gaussian and T methods, which can be parallelized across tests. The number of clusters, K, increases like $\log(T)$ (Rodriguez and Müller 2013, p. 27), so the DPM sampler scales like $T\log(T)$ in the number of tests. The DPM sampler cannot be fully parallelized, but our approach could be optimized via implementation in a low-level language. It could also be split into multiple MCMC chains to reduce the required length of each individual chain. Previous model fits could also be used to "warm start" the sampler, removing the need for a long warm-up period.

Because the samplers involve $p \times p$ matrix inversions, they scale as p^3 in the number of metrics. Consequently, including a large number of business metrics (say, $p > 20$) could prove computationally prohibitive in the absence of programming optimizations. In most cases, we believe that our framework could be successfully applied to settings with $p \le 10$, especially if researchers use the surrogacy measures of Section 4.1 to remove low-quality proxy metrics. Consequently, adjustments to the computational approach are likely needed only in exceptional cases. If computation time is a concern, practitioners could leverage work on scalable Bayesian computational methods, such as variational inference (Blei and Jordan 2006) or distributed MCMC (Bumbaca et al. 2020).

When applied in practice, the meta-analysis methods use a large database of previous experiments to generate inferences for a much smaller set of executing experiments. Thus, one simple strategy to reduce computational demands is to refit the meta-analysis method in batches (e.g., weekly) and apply the most-recent posterior samples of $P_{\boldsymbol{\theta}}$ to generate estimates for currently running tests. This approach is only approximate in that it ignores recent data when estimating $P_{\boldsymbol{\theta}}$, but with a large number of previous experiments, the approximation is likely to be quite accurate. This batch estimation method is particularly straightforward for the Gaussian and T methods. It is also possible to apply it to the DPM method, although doing so would require a different model-fitting strategy that produces explicit estimates of $P_{\boldsymbol{\theta}}$.

Appendix C. Sensitivity Analysis

We performed a sensitivity analysis that mimics the sample-split analysis of Section 6.2, substituting a different prior distribution that meaningfully changes almost every hyperparameter. In particular, we set $\boldsymbol{\Phi} = 4\mathbf{I}_p$, $\nu = 16$, and $\boldsymbol{\Psi} = 64\mathbf{I}_p$ for the Gaussian and T methods; we did not change the value of $\boldsymbol{\eta}_0$ ($= \mathbf{0}$) because this is the most natural choice for this hyperparameter. For the T method, we assumed the prior $\log(\gamma) \sim \mathcal{N}(1, 0.5)$ with a prior median of $e \approx 2.7$ degrees of freedom. For the DPM, we set hyperpriors as follows: $\boldsymbol{\eta}_0 \sim \mathcal{N}(\mathbf{0}_p, 0.01\mathbf{I}_p)$, $\boldsymbol{\Psi} \sim \mathcal{W}^{-1}(48\mathbf{I}_p, 12)$, $\phi \sim \mathcal{G}(3,1)$, $\alpha_0 \sim \mathcal{G}(1,2)$, and $\nu = 20$.

Employing this prior distribution resulted in only small changes to the results of Section 6.2. Table C.1 displays results analogous to Table 3 under this new prior distribution. The results indicate that the meta-analytic methods still substantially outperform the naive approach, and the performance ordering is similar to that of Table 3: DPM, T, Gaussian-0, Gaussian, and naive (from best to worst). None of the other results substantially alter any conclusions from Section 6.2. However, the lower prior mean for α_0 results in a smaller estimated number of clusters.

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Table C.1. Pair-Wise Method Comparisons for Predictive Performance in Sensitivity Analysis

Metric	Method	Naive	Gaussian-0	Gaussian	T	DPM
Win Rate	Naive	—	0.27 (0.03)	0.29 (0.03)	0.27 (0.03)	0.32 (0.03)
	Gaussian-0	0.73 (0.03)	—	0.56 (0.04)	0.31 (0.03)	0.41 (0.04)
	Gaussian	0.71 (0.03)	0.44 (0.04)	—	0.33 (0.03)	0.41 (0.04)
	T	0.73 (0.03)	0.69 (0.03)	0.67 (0.03)	—	0.38 (0.04)
	DPM	0.68 (0.03)	0.60 (0.04)	0.59 (0.04)	0.62 (0.04)	—
Δ*Mean* log *Error Quotient*	Naive	—	0.10 (0.05)	0.09 (0.05)	0.16 (0.06)	0.30 (0.08)
	Gaussian-0	−0.10 (0.05)	—	0.00 (0.00)	0.07 (0.03)	0.21 (0.07)
	Gaussian	−0.09 (0.05)	0.00 (0.00)	—	0.07 (0.03)	0.21 (0.07)
	T	−0.16 (0.06)	−0.07 (0.03)	−0.07 (0.03)	—	0.14 (0.06)
	DPM	−0.30 (0.08)	−0.21 (0.07)	−0.21 (0.07)	−0.14 (0.06)	—

Notes. Pair-wise predictive performances of five methods on the sensitivity analyses are shown with approximate standard errors in parentheses; compare with Table 3. The *Win Rate* section displays the proportion of tests in which the method listed in the row outperformed the method listed in the column. In the Δ*Mean* log *Error Quotient* section, a negative (positive) value indicates that the row (column) method outperformed the column (row) method.

Endnotes

[1] Yang et al. (2024) show that this assumption can be relaxed somewhat to require only that the sign of the conditional average treatment effect (CATE) matches that of a surrogate-imputed CATE.

[2] To be more precise, we could add an additional subscript to m to acknowledge that these outcomes are defined for individual customers. Our notation omits this dependence for brevity.

[3] Rather than treating the analysis as a multivariate problem, practitioners often define one combined metric as a linear combination of individual metrics; this approach is easily seen to be equivalent to the naive approach owing to the linearity of both the estimator, $\mathbf{z}_t$, and the combined metric.

[4] In Bayesian modeling, a hyperprior refers to a prior distribution for a prior distribution. The parameters of the hyperprior are often referred to as hyperparameters.

[5] These could be extended to the metaregression setting by introducing a mixed-effects structure, such as $\boldsymbol{\mu}_t = \boldsymbol{\alpha}_t + \mathbf{B}\mathbf{w}_t$, where $\boldsymbol{\alpha}_t \in \mathbb{R}^p$ is a mean-zero random effect and $\mathbf{B} \in \mathbb{R}^{d\times s}$ is a matrix of fixed-effect regression coefficients. A further generalization might allow the distribution of $\boldsymbol{\alpha}_t$ to depend on $\mathbf{w}_t$. These applied modeling techniques are largely context driven and have been used frequently in other areas of quantitative marketing; thus, we aim to provide only an introduction to the basic modeling techniques and leave the details of these generalizations to future applied work.

[6] Specifically, Assumption 4 must continue to hold given missingness indicator variables. If the missingness is believed to be systematic, analysts can handle this challenge by specifying a metaregression model in which the distribution of $\boldsymbol{\mu}_t$ depends on the missingness indicators.

[7] In principle, the effect of internal surrogacy could be removed from PVR_t by replacing $\boldsymbol{\Omega}_{\mathbf{yy}}$ with $\text{diag}(\boldsymbol{\Omega}_{\mathbf{yy}})$ in the definition of $v_{\mathbf{y}t}$, corresponding to an analysis in which we shrink the estimates $\mathbf{y}_t$ toward their prior mean ignoring their prior correlations. Because this analysis is less natural than the analysis that simply discards $\mathbf{x}_t$, we have opted to define $v_{\mathbf{y}t}$ (and consequently, PVR_t and SVR_t) using $\boldsymbol{\Omega}_{\mathbf{yy}}$.

[8] To make the simulation more realistic, we did not attempt to set prior parameters to closely match those of the data-generating process (from the fitted DPM model). In general, Bayesian methods do not deliver exact coverage under prior misspecification in finite samples.

[9] We omit the differences in coverage rates because these differences all entail a standard error that rounds to 0.00.

[10] Specifically, we used agglomerative clustering with complete linkage. We specified the distance metric as the posterior probability that a pair of tests belongs to different clusters.

References

Ansari A, Mela CF (2003) E-customization. *J. Marketing Res.* 40(2):131–145.

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *J. Marketing Res.* 55(1):80–98.

Athey S, Chetty R, Imbens GW, Kang H (2026) The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *Rev. Econom. Stud.* 93(4):2284–2312.

Azevedo EM, Deng A, Montiel Olea JL, Weyl EG (2019) Empirical Bayes estimation of treatment effects with many A/B tests: An overview. *AEA Papers Proc.* 109:43–47.

Azevedo EM, Deng A, Montiel Olea JL, Rao J, Weyl EG (2020) A/B testing with fat tails. *J. Political Econom.* 128(12):4614–4672.

Berger JO (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics, 2nd ed. (Springer-Verlag, New York).

Berman R (2018) Beyond the last touch: Attribution in online advertising. *Marketing Sci.* 37(5):771–792.

Blei DM, Jordan MI (2006) Variational inference for Dirichlet process mixtures. *Bayesian Anal.* 1(1):121–143.

Brooks S, Gelman A, Jones G, Meng XL (2011) *Handbook of Markov Chain Monte Carlo* (CRC Press, Boca Raton, FL).

Brown LD, Purves R (1973) Measurable selections of extrema. *Ann. Statist.* 1(5):902–912.

Bumbaca F, Misra S, Rossi PE (2020) Scalable target marketing: Distributed Markov chain Monte Carlo for Bayesian hierarchical models. *J. Marketing Res.* 57(6):999–1018.

Burr D, Doss H (2005) A Bayesian semiparametric model for random-effects meta-analysis. *J. Amer. Statist. Assoc.* 100(469):242–251.

Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H (2000) The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 1(1):49–67.

Conlon A, Taylor J, Li Y, Diaz-Ordaz K, Elliott M (2017) Links between causal effects and causal association for surrogacy evaluation in a Gaussian setting. *Statist. Medicine* 36(27):4243–4265.

Cornish EA (1954) The multivariate *t*-distribution associated with a set of normal sample deviates. *Australian J. Phys.* 7(4):531–542.

Deng A (2015) Objective Bayesian two sample hypothesis testing for online controlled experiments. Gangemi A, Leonardi S, Panconesi A (General Chairs), eds. *Proc. 24th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 923–928.

Deng A, Shi X (2016) Data-driven metric development for online controlled experiments: Seven lessons learned. Krishnapuram B, Shah M (General Chairs), Smola A, Aggarwal C, Shen D, Rastogi R (Program Chairs), eds. *Proc. 22nd ACM SIGKDD*

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

Internat. Conf. Knowledge Discovery Data Mining (Association for Computing Machinery, New York), 77–86.

Deng A, Xu Y, Kohavi R, Walker T (2013) Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. Leonardi S, Panconesi A (General Chairs), Ferragina P, Gionis A (Program Chairs), eds. *Proc. Sixth ACM Internat. Conf. Web Search Data Mining* (Association for Computing Machinery, New York), 123–132.

Deng A, Hagar L, Stevens N, Xifara T, Yuan LH, Gandhi A (2023) From augmentation to decomposition: A new look at CUPED in 2023. Preprint, submitted December 5, https://doi.org/10.48550/arXiv.2312.02935.

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials* 7(3):177–188.

DerSimonian R, Laird N (2015) Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials* 45(Pt A):139–145.

Dew R, Ansari A (2018) Bayesian nonparametric customer base analysis with model-based visualizations. *Marketing Sci.* 37(2):216–235.

Ding P (2016) On the conditional distribution of the multivariate *t* distribution. *Amer. Statistician* 70(3):293–295.

Efron B (2014) Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* 29(2):285–301.

Elliott MR (2023) Surrogate endpoints in clinical trials. *Annual Rev. Statist. Appl.* 10(1):75–96.

Elliott MR, Conlon AS, Li Y, Kaciroti N, Taylor JM (2015) Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics* 16(2):400–412.

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* 90(430):577–588.

Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, Bauer S, Kilbertus N, Kohane IS, van der Schaar M (2024) Causal machine learning for predicting treatment outcomes. *Nature Medicine* 30(4):958–968.

Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6(4):733–760.

Hair JF Jr, Sarstedt M (2021) Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing. *J. Marketing Theory Practice* 29(1):65–77.

Henderson CR (1950) Estimation of genetic parameters (abstract). *Ann. Math. Statist.* 21(2):309–310.

Higgins JP, Thompson SG, Spiegelhalter DJ (2009) A re-evaluation of random-effects meta-analysis. *J. Roy Statist. Soc. Ser. A Statist. Soc.* 172(1):137–159.

Huang S, Wang C, Yuan Y, Zhao J, Zhang BJ (2026) Estimating effects of long-term treatments. *Management Sci.*, ePub ahead of print April 9, https://doi.org/10.1287/mnsc.2023.02575.

Imbens G, Kallus N, Mao X, Wang Y (2024) Long-term causal inference under persistent confounding via data combination. *J. Roy Statist. Soc. Ser. B Statist. Methodology* 87(2):362–388.

Jeunen O, Ustimenko A (2024) Learning metrics that maximise power for accelerated A/B-tests. Baeza-Yates R, Bonchi F (General Chairs), eds. *Proc. 30th ACM SIGKDD Conf. Knowledge Discovery Data Mining KDD 24* (Association for Computing Machinery, New York), 5183–5193.

Jin Y, Ba S (2023) Toward optimal variance reduction in online controlled experiments. *Technometrics* 65(2):231–242.

Joffe MM, Greene T (2009) Related causal frameworks for surrogate outcomes. *Biometrics* 65(2):530–538.

Kim JG, Menzefricke U, Feinberg FM (2004) Assessing heterogeneity in discrete choice models using a Dirichlet process prior. *Rev. Marketing Sci.* 2(1):1–41.

Koenker R, Mizera I (2014) Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* 109(506):674–685.

Kohavi R, Tang D, Xu Y (2020) *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*, 1st ed. (Cambridge University Press, Cambridge, UK).

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. *Quart. J. Econom.* 130(4):1941–1973.

Liu JS, Wong WH, Kong A (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* 81(1):27–40.

Meng XL (1994) Posterior predictive *p*-values. *Ann. Statist.* 22(3):1142–1160.

Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: An overview. *WIREs Data Mining Knowledge Discovery* 2(1):86–97.

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* 9(2):249–265.

Neal RM (2011) MCMC using Hamiltonian dynamics. Brooks S, Gelman A, Jones GL, Meng XL, eds. *Handbook of Markov Chain Monte Carlo* (CRC Press, Boca Raton, FL), 113–162.

Pan SJ, Yang Q (2009) A survey on transfer learning. *IEEE Trans. Knowledge Data Engrg.* 22(10):1345–1359.

Rafieian O (2023) A matrix completion solution to the problem of ignoring the ignorability assumption. Preprint, submitted September 6, https://dx.doi.org/10.2139/ssrn.4546996.

Ray K, van der Vaart A (2021) On the Bernstein–von Mises theorem for the Dirichlet process. *Electronic J. Statist.* 15(1):2224–2246.

Robinson GK (1991) That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* 6(1):15–32.

Rodriguez A, Müller P (2013) Nonparametric Bayesian inference. *NSF–CBMS Regional Conf. Ser. Probab. Statist.*, vol. 9, i–110.

Rossi PE, Allenby GM, Misra S (2024) *Bayesian Statistics and Marketing*, Wiley Series in Probability and Statistics, 2nd ed. (John Wiley & Sons, Newark, NJ).

Rousseau J (2016) On the frequentist properties of Bayesian nonparametric methods. *Annual Rev. Statist. Appl.* 3(1):211–231.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* 66(5):688–701.

Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* 12(4):1151–1172.

Sethuraman J, Ghosh M (2024) Some properties of GEM(α) distributions. *Sankhya A* 86(Suppl 1):288–300.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. *J. Roy Statist. Soc. Ser. B Statist. Methodology* 64(4):583–639.

Srinivasan S, Vanhuele M, Pauwels K (2010) Mind-set metrics in market response models: An integrative approach. *J. Marketing Res.* 47(4):672–684.

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* 82(398):528–540.

Tripuraneni N, Richardson L, D'Amour A, Soriano J, Yadlowsky S (2024) Choosing a proxy metric from past experiments. Baeza-Yates R, Bonchi F (General Chairs), eds. *Proc. 30th ACM SIGKDD Conf. Knowledge Discovery Data Mining KDD 24* (Association for Computing Machinery, New York), 5803–5812.

van Dyk DA, Meng X-L (2001) The art of data augmentation (with discussion). *J. Comput. Graph. Statist.* 10(1):1–50.

van Houwelingen HC, Arends LR, Stijnen T (2002) Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statist. Medicine* 21(4):589–624.

Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Machine Learn. Res.* 11(116):3571–3594.

Yang J, Eckles D, Dhillon P, Aral S (2024) Targeting for long-term outcomes. *Management Sci.* 70(6):3841–3855.

Easton Huch is a postdoctoral fellow in marketing at the Johns Hopkins Carey Business School. His research addresses decision-making problems in digital marketing platforms, customer relationship management, product development, and advertising. While working on this article, he was employed as an applied scientist intern at Zillow. During this time, he also completed his PhD in statistics at

Downloaded from informs.org by [216.73.217.10] on 05 November 2026, at 15:00 . For personal use only, all rights reserved.

the University of Michigan under the supervision of Walter H. Dempsey and Fred Feinberg.

Stephanie A. Nail is Manager, Data Science, Experimentation Platform at Zillow, after joining as an applied scientist. Her work bridges rigorous quantitative research and real-world, data-driven decisions across experimentation and causal inference. She previously worked on Meta's Bayesian Modeling team, holds a PhD in Political Science, and completed a postdoctoral research fellowship at Stanford. Her work appears in *Research in Politics* and the *Oxford Encyclopedia of Political Decision Making*.

Terran Melconian is a principal applied scientist at Zillow. Their work focuses on evaluation, measurement, and decision making under uncertainty, including A/B testing and evaluations of generative artificial intelligence products. They have degrees in aeronautical and astronautical engineering from the Massachusetts Institute of Technology and previously worked on machine learning and ranking problems at Google, TripAdvisor, and Meta.

Xan Vongsathorn is a principal applied scientist at Zillow, where he specializes in building measurement systems, including Zillow's A/B testing platform and systems for offline evaluation of machine learning models and generative artificial intelligence products. He has a PhD in economics from the University of Chicago with a focus on the design of online marketplaces. He previously worked on machine learning, causal inference, and A/B testing at Microsoft.

Fred Feinberg is Handleman Professor of Marketing at the Ross School of Business, and Professor of Statistics, University of Michigan. He received his PhD at MIT-Sloan under John D. C. Little, and his research focuses on choice models, Bayesian methodology, nonparametrics, causal inference, and the marketing/engineering design interface. He is a past president of *INFORMS Society for Marketing Science*, and former coeditor at *Marketing Science* and *Production and Operations Management*.