

Marketing Science Institute Working Paper Series 2026

Report No. 26-115

# Predicted Incrementality by Experimentation (PIE) for Ad Measurement

Brett R. Gordon, Roberta Moakler and Florian Zettelmeyer

"Predicted Incrementality by Experimentation (PIE) for Ad Measurement" © 2026

Brett R. Gordon, Roberta Moakler and Florian Zettelmeyer

MSI Working Papers are Distributed for the benefit of MSI corporate and academic members and the general public. Reports are not to be reproduced or published in any form or by any means, electronic or mechanical, without written permission.

# Predicted Incrementality by Experimentation (PIE) for Ad Measurement*

Brett R. Gordon
Kellogg School of Management
Northwestern University

Robert Moakler
Ads Research
Meta

Florian Zettelmeyer
Kellogg School of Management
Northwestern University and NBER

April 1, 2026

**Abstract**

Randomized controlled trials (RCTs) provide the most credible estimates of advertising incrementality but are difficult to scale. We propose Predicted Incrementality by Experimentation (PIE), which reframes ad measurement as a campaign-level prediction problem. PIE uses a sample of RCTs to learn a mapping from campaign features to causal effects, then applies it to campaigns not run as RCTs. Because the RCTs identify the causal effects, PIE can incorporate post-determined features—campaign-level aggregates such as test-group outcomes, exposure rates, and last-click conversions, computed after campaign completion. These metrics reflect the consumer behaviors that generate treatment effects, so they carry predictive information about incrementality even though they would be invalid controls in a causal model. Using 2,226 Meta ad experiments, PIE achieves an out-of-sample $R^2 = 0.88$ for incremental conversions per dollar, compared to $R^2 = 0.19$ for industry-standard 7-day last-click attribution. In a decision-making framework, PIE disagrees with RCT-based decisions in only 8–12% of campaigns, compared to 12–20% for last-click attribution. We conclude that PIE can help scale causal measurement from a limited number of RCTs to a large set of non-experimental campaigns.

**Keywords:** Digital Advertising, Advertising Effects, Advertising Measurement, Field Experiments, RCTs, Causal Inference, Attribution Metrics.

* We thank Neha Bhargava, JP Dubé, Dean Eckles, Elea Feit, Garrett Johnson, Randall Lewis, Jim Lecinski, Oded Netzer, Ilya Morozov, Kelly Paulson, Julian Runge, Ryan Shyu, Andrey Simonov, and seminar participants at Arizona State University, Carnegie Mellon University, Columbia University, Cornell University, Georgetown, HKUST, Invitational Choice Symposium, MIT, Notre Dame, Purdue University, OSU, Stanford, UNC, University of Arizona, University of Iowa, University of Maryland, University of Michigan, University of Minnesota, University of Rochester, UT Austin, Virtual Quantitative Marketing Seminar, Washington University in St. Louis, the Utah Winter Business Economics Conference, the Conference on Digital Experimentation, the Marketing Science Conference, the NYU-Temple-CMU Conference, Amazon, Meta, and MSI. Previously, Gordon and Zettelmeyer were part-time employees of Facebook (Academic Researchers, 3 hours/week). Currently, Gordon and Zettelmeyer hold concurrent appointments at Northwestern and Amazon. This paper describes work performed at Northwestern and is not associated with Amazon. E-mail addresses for correspondence: b-gordon@kellogg.northwestern.edu, f-zettelmeyer@kellogg.northwestern.edu, rmoakler@meta.com.

# 1 Introduction

Marketers are under pressure to demonstrate that digital advertising delivers benefits. The relevant causal quantity is "incrementality"—a comparison of outcomes under an ad campaign to the counterfactual outcomes without it. Randomized controlled trials (RCTs) offer the most credible approach to measuring causal effects, but they impose substantial burdens: beyond engineering and operational complexity in auction markets, withholding ad exposures from some users creates opportunity costs for both advertisers and platforms. As a result, advertisers are reluctant to run RCTs across all campaigns (e.g., Johnson, 2023). In practice, many rely instead on attributed metrics, such as last-click conversions, or on observational methods that exploit quasi-experimental variation in advertising exposure. While these approaches scale well, both can systematically bias causal ad effects when exposure and outcomes are jointly driven by user intent and algorithmic targeting (Lewis et al., 2011; Lewis and Reiley, 2014; Lewis and Rao, 2015; Gordon et al., 2019, 2023).[1]

We propose Predicted Incrementality by Experimentation (PIE), a method that uses a sample of RCTs to predict treatment effects for non-experimental campaigns. Predictors can include both pre-determined features known before the campaign runs (e.g., objective, budget, vertical) and post-determined features observed after the campaign concludes (e.g., average outcomes, exposure rate, click-through rate, last-click conversions). Post-determined features carry predictive information about incrementality because they reflect the same consumer behaviors and selection processes that generate treatment effects, even though they are not themselves causal quantities. The RCTs provide unbiased estimates of causal effects, and PIE learns how those effects covary with campaign-level observables. Because identification is handled by the experiments, endogenous campaign metrics that would be inappropriate controls in a causal model (e.g., last-click conversions) can serve as informative predictors of the treatment effects. These features remain available outside of RCTs because they can be computed from the treated group in experiments and from delivered campaign data in non-experimental settings.[2] In this sense, PIE is a prediction problem rather than a causal identification problem.

We formalize this intuition using a series of stylized models in Section 3. We begin by motivating PIE through the fundamental problem of causal inference: for each user, potential outcomes under treatment and control cannot both be observed, so causal inference can be viewed as a missing data problem (Imbens and Rubin, 2015). Traditional approaches address this problem either by estimating average effects directly, for example with difference-in-means or inverse probability weighting, or by imputing unobserved counterfactuals under modeling assumptions.

We adopt this lens but shift the unit of analysis from the user to the campaign. For RCT

[1] Last-click conversions attribute a conversion if the most recent ad was clicked within a predefined attribution window (e.g., 7 or 14 days). If no ad is clicked, no conversion is attributed to the campaign via click-through attribution. These counts can overestimate the causal effect of ads if ads are targeted to users who would have purchased anyway, or underestimate it if impressions that drive conversions fall outside the attribution window.

[2] Post-determined features and treatment effect estimates can share mechanical sampling variation because both are computed from the same experimental sample. We discuss this concern and why it is unlikely to materially affect our results in Section 4.2.

campaigns, we observe average outcomes under both treatment and control; for non-experimental campaigns, we observe only the average outcome in the treated population. As a starting point, we consider the simplest version of PIE, which assumes that all treated consumers are exposed to the ad (perfect compliance in exposure). Under this assumption, the model estimates a mapping between average control and test outcomes across RCTs and applies that mapping to impute the missing average control outcome for a non-experimental campaign. This yields out-of-sample predictions of causal effects for campaigns without a randomized control group.

We next generalize the model to allow for one-sided noncompliance in ad exposure and for intermediate behaviors such as ad clicks that may causally affect the outcome. In this richer setting, post-determined features remain informative predictors because they reflect the same underlying factors that shape treatment effects, even when they are endogenous to treatment. More generally, any correlation between campaign-level observables and treatment effects can aid prediction, regardless of whether that correlation arises from causal or confounded relationships.[3] In particular, unobserved factors that simultaneously drive ad exposure, intermediate behaviors, and consumer outcomes induce cross-campaign correlations among observable metrics that a predictive model can exploit.

This logic also applies to intermediate behaviors. Even if clicking on an ad has no direct causal effect on conversions, the campaign-level click-through rate can still predict treatment effects if click decisions and conversion outcomes are shaped by some of the same unobservable factors. More generally, post-determined features can remain informative predictors even when they are endogenous, because they reflect the heterogeneity that drives variation in treatment effects across campaigns.

The strength of this predictive power depends on the underlying data generating process. We use simulations to characterize when PIE should perform better or worse relative to a benchmark that predicts the same average treatment effect for every campaign. The simulations identify six conditions under which PIE performs better, including greater heterogeneity in treatment effects, a stronger positive correlation between baseline outcomes and treatment effects, and stronger selection of consumers into exposure.

In our application, we implement PIE using a random forest trained on a dataset of 2,226 Meta ad experiments, illustrating how the approach can be operationalized with a flexible supervised learning model. We use this application both to demonstrate feasibility and to show how predictive accuracy varies across campaign types and segments. The overall PIE model achieves an out-of-sample $R^2$ of 0.88 for incremental conversions per dollar spent, so predictive accuracy reflects cross-campaign heterogeneity rather than variation in campaign scale. The model performs comparably across key subgroups such as conversion funnel levels (lower, mid, upper), prospecting versus retargeting campaigns, and industry verticals. PIE also achieves higher out-of-sample $R^2$ than last-click conversions across these subgroups. Its performance improves substantially as the

[3] In a standard causal inference setting, unobserved confounders that jointly affect treatment and outcomes would violate the backdoor criterion (Pearl, 2009) and bias causal estimates. For PIE, these same confounders are a *source* of predictive power rather than a threat to identification.

number of available training RCTs increases. Training with 50 RCTs yields $R^2 = 0.37$, increasing to $R^2 = 0.72$ with 400 RCTs and $R^2 = 0.81$ with 1,600 RCTs, approaching the full-sample performance of $R^2 = 0.88$.

We investigate how accuracy degrades when PIE extrapolates. The model's predictions generalize well to new campaigns by the same advertiser ($R^2 = 0.87$) but less well to advertisers absent from the training data ($R^2 = 0.72$). We examine model performance when a particular segment of advertisers is absent from the training data. Across six segmentation variables—including advertiser size, vertical, campaign year, and audience type—PIE predicts less accurately when extrapolating than when training includes the focal segment, but the degree of degradation varies substantially. For example, extrapolating from retargeting to prospecting campaigns reduces $R^2$ by only 0.9 percentage points, whereas extrapolating to travel advertisers from non-travel advertisers reduces $R^2$ by 25 percentage points and extrapolating across campaign years reduces it by 21.3 percentage points.

Finally, the purpose of measurement is ultimately to inform decisions. Improved prediction accuracy has limited practical value unless it changes decisions. To create a decision-making framework, we adopt a simple perspective drawn from industry practice in which marketers consider a campaign successful if its performance exceeds some threshold; otherwise, the campaign is viewed as unsuccessful, and the marketer would have preferred to allocate the campaign budget to another business activity with greater expected returns. We use this framing to quantify the disagreement probability between decisions made using RCTs and those using predictive models. Specifically, we quantify how often decisions based on PIE or last-click would disagree with decisions based on RCTs. We find that PIE exhibits low disagreement rates and a better balance of Type I and Type II errors than last-click attribution, indicating that even when prediction errors remain, managers using PIE would more often reach the same go/no-go decisions as with RCTs.

Methodologically, we make two related contributions. First, we reframe non-RCT measurement as a campaign-level prediction problem (as opposed to a traditional causal inference problem), treating each RCT as a labeled observation and learning a mapping from campaign features to causal effects. Second, we introduce post-determined features and formalize the conditions under which they carry information about incrementality, showing that variables often dismissed as biased proxies for causal effects can be valid and informative predictors. To the best of our knowledge, we are the first to propose the PIE approach. Although our focus is on using PIE for ad measurement, the methodology could be generalized to other types of marketing interventions.[4]

Empirically and managerially, we offer a proof-of-concept using a large dataset of campaigns from a major advertising platform. PIE achieves higher predictive accuracy than common attribution benchmarks. We also propose a decision-theoretic framework for evaluating model-based incremental performance by focusing on the decision errors managers would make using these approaches relative to an RCT.

[4]Our approach differs from the industry "calibration factor" practice that scales attributed metrics by a single ratio from one RCT and reuses it thereafter (Google, 2024, p. 15). Our formalization explains when and why this mapping works, generalizing beyond a single scaling factor by learning campaign-specific corrections.

Our approach has been adopted in industry practice. An implementation of our model is in production at Amazon Advertising, where it is used to calibrate machine learning attribution models—composed of post-determined features—to RCT outcomes, enabling the attribution models to assign touchpoint-level credit in Amazon's new multi-touch attribution (MTA) system (Lewis et al., 2025).[5] Meta's Incremental Attribution makes use of similar approaches to the PIE models described here—leveraging pre- and post-determined features across a large set of prior RCTs. This model was developed independently by Meta during the same period as our research.[6]

We caution against generalizing our specific empirical results to other prediction settings (e.g., different platforms). Instead, we view these results as a proof-of-concept demonstrating that the approach can work using actual RCT data from a major advertising platform. In practice, an ad platform planning to implement PIE would need to conduct out-of-sample validation tests to determine whether PIE performs sufficiently well for its intended purpose. As with any prediction problem, our approach relies on having a representative training sample and assumes the mapping between features and outcomes is stable. Platforms can design their own RCT data generation process to help ensure that these conditions are met.

The remainder of the paper proceeds as follows. Section 2 discusses our work's relationship to several literatures. Section 3 develops the formal model and sources of predictive power. Section 4 introduces the application of PIE to ad experiments at Meta, and Section 5 presents results and robustness checks, including subgroup analyses and tests of generalizability. Section 6 evaluates alignment with managerial decisions. Section 7 offers concluding remarks and guidance for deploying and maintaining PIE at platform scale.

## 2 Relationship to Existing Literature

Our work relates to five distinct areas of research: advertising effectiveness, methods that combine experimental and observational data, causal machine learning and uplift modeling, transportability and external validity, and meta-analysis. We discuss each in turn, emphasizing how PIE differs where appropriate.

**Advertising Effectiveness.** Our work builds on the literature measuring advertising effectiveness; see Gordon et al. (2021) and Johnson (2023) for recent reviews. This literature finds that ad effects are typically small and that detecting them requires careful experimental design and large samples (Lewis et al., 2011; Lewis and Reiley, 2014; Lewis and Rao, 2015). In the absence of RCTs, researchers have exploited quasi-experimental variation in ad auctions (Rafieian and Yoganarasimhan, 2022; Gui et al., 2022), and platforms have improved experimental practice through eligibility-based designs such as ghost ads (Johnson et al., 2017). These advances are important because even high-quality observational data, combined with sophisticated machine learning methods, can fail to recover incrementality when exposure and outcomes are jointly determined by user

[5] See also the announcement at Amazon unBoxed 2024, `https://tinyurl.com/3wetscdm`.
[6] See slide 16 of `https://tinyurl.com/2kmuwj3n`.

intent and targeting (e.g., Gordon et al., 2019, 2023). At the same time, conducting RCTs for all campaigns is resource intensive: beyond engineering complexity, withholding ad exposure (or impression opportunities) imposes opportunity costs that make continuous testing difficult to scale for many advertisers.

Against this backdrop, PIE reframes measurement as a campaign-level prediction problem trained on experimental data. Rather than modeling user-level counterfactuals, we treat each RCT as a labeled observation and learn a mapping from campaign features—both pre-determined configuration settings and post-determined metrics—to causal ad effects. We then apply this learned mapping to non-experimental campaigns without requiring control group data in production. Variables that would normally be invalid as post-treatment controls in user-level causal models (e.g., last-click conversions) can be valid predictors in a campaign-level prediction task (see Angrist and Pischke, 2009; Hernán and Robins, 2020, on "bad controls" and post-treatment variables).

**Combining Experimental and Observational Data.** PIE is related to work that blends experimental and observational data, though it differs in its target and granularity. One stream of research calibrates marketing mix models (MMMs) to lift tests, using experiments as priors or targets in Bayesian MMMs (Runge et al., 2023, 2025; Jin et al., 2017). Another uses surrogate outcomes to link short-run metrics to long-run targets via experiments (Athey et al., 2025; Yang et al., 2024). PIE shares this experiment-anchored approach but applies it to campaign-level causal prediction rather than channel-level calibration or long-run forecasting. Amazon's multi-touch attribution system provides another related example, with PIE extrapolating RCT effects across campaigns before the system allocates attribution credit to individual touchpoints (Lewis et al., 2025).

**Causal Machine Learning and Uplift Modeling.** PIE relates to causal machine learning and marketing uplift modeling. Methods in this area use experimental data to estimate heterogeneous treatment effects at the individual level and to rank or target users by predicted incremental response (Athey and Imbens, 2016; Wager and Athey, 2018; Rzepakowski and Jaroszewicz, 2012; Ascarza, 2018). PIE shares the logic of using experiments to supply labels and machine learning to generalize, but differs in unit of analysis (campaigns rather than users) and objective (aggregate effects for measurement rather than individual effects for targeting). Related in spirit is Huang et al. (2024), who pool multiple experiments to predict individual-level treatment effects for personalization. PIE differs in targeting campaign-level effects and in leveraging post-determined features for prediction, aiming for accurate out-of-sample prediction rather than causal explanation (Shmueli, 2010).

**External Validity and Transportability.** PIE draws on the literature on external validity and transportability (Pearl and Bareinboim, 2011). Conceptually, we treat the RCT dataset as the source domain and non-experimental campaigns as the target domain, and we assume that the mapping from campaign features to causal effects is stable across these domains. This is

an invariance assumption: the conditional relationship $f(\mathbf{X}_r;\theta)$ estimated on RCT campaigns generalizes to non-RCT campaigns with similar feature distributions. We assess this assumption through out-of-sample validation. The economics literature has emphasized related challenges of external validity (Allcott, 2015; Vivalt, 2020), and Hotz et al. (2005) use observed covariates to predict treatment efficacy in new settings, which is conceptually close to PIE's use of campaign features to predict treatment effects for new campaigns. PIE also has a connection to transfer learning, in which a model trained on one task or domain is adapted to another. However, transfer learning typically involves a shift in the prediction task itself, whereas PIE performs the same task (predicting campaign incrementality) in both the training and prediction stages.

**Meta-Analysis.** Finally, PIE can be viewed as a predictive, study-level extension of meta-analysis. Like random-effects meta-analysis, it pools results from many independent experiments while allowing for between-study heterogeneity. Traditional meta-analysis or meta-regression typically targets the grand mean effect and its dispersion or estimates a small set of moderator coefficients (Stanley and Doucouliagos, 2012). PIE departs from this tradition by aiming to generate accurate out-of-sample *point predictions* for individual new campaigns, leveraging the full heterogeneity in the RCT sample to map campaign-specific features to campaign-specific causal effects.

# 3 Model of Predicted Incrementality by Experimentation

We introduce our methodology from the perspective of an advertising platform with access to a large and representative sample of RCTs. The platform seeks to provide estimates of incremental ad effects for campaigns that were not implemented as RCTs. PIE uses the RCT dataset to train a predictive model at the campaign level that enables calculation of incremental effects for these non-experimental campaigns.

Section 3.1 begins with model preliminaries that specify user-level behavior in the context of a collection of ad campaigns and relate these quantities to their aggregated campaign-level counterparts. In Section 3.2, we introduce the PIE approach under the assumption of perfect compliance in treatment assignment, meaning that all users in the test group are exposed to ads while no users are exposed in the control group. This assumption allows us to communicate the intuition and mechanics of the PIE model. Section 3.3 generalizes the model to allow for one-sided imperfect compliance with selection into exposure, which provides a more realistic description of digital ad experiments. Section 3.4 presents simulation results that characterize the conditions under which the PIE model performs better or worse. Section 3.5 generalizes the PIE model from the potential outcomes framing in Sections 3.2 and 3.3 to an arbitrary supervised learning model $f(\mathbf{X}_r;\theta)$ that maps a vector of pre- and post-determined campaign features to the causal target.

## 3.1 Model Preliminaries

We observe a collection of $R$ independent campaigns indexed by $r = 1, \ldots, R$ that were implemented as RCTs. Let $\mathcal{R} = \{1, \ldots, R\}$ denote this set of RCT campaigns. We make the standard

assumptions of no interference between units (SUTVA) and no spillover effects across campaigns.

Within each RCT $r$, users are indexed by $i = 1, \ldots, N_r$ and are randomly assigned to either a test group (denoted by $Z_i^r = 1$) or a control group ($Z_i^r = 0$) with sizes $N_{tr}$ and $N_{cr}$, respectively.

Assignment to the test group enables the possibility of ad exposure, while control group users are prevented from seeing the ad. We denote actual exposure by $D_i^r = 1$ (exposed) and $D_i^r = 0$ (not exposed). In general, $D_i^r$ and $Z_i^r$ may differ due to one-sided noncompliance, which we address in Section 3.3.

Each RCT aims to measure the impact of the campaign on the observed outcome metric $Y_i^r \in \mathbb{R}$, such as purchasing a product from the advertiser. We adopt the potential outcomes framework such that $Y_i^r(d)$ denotes the potential outcome for user $i$ in RCT $r$ under exposure status $d \in \{0, 1\}$.

A user's baseline outcome (without exposure) is $\alpha_i^r = Y_i^r(0)$, and the user-specific treatment effect is $\tau_i^r = Y_i^r(1) - Y_i^r(0)$. The observed outcome is thus

$$Y_i^r = Y_i^r(D_i^r) = \alpha_i^r + \tau_i^r D_i^r. \tag{1}$$

For each RCT $r$, let $\tau_r$ and $\alpha_r$ denote the campaign-level average treatment effect (ATE) and average baseline outcome, respectively. Across campaigns these parameters are heterogeneous, drawn respectively from distributions $G_\tau$ and $G_\alpha$:

$$\tau_r = \mathbb{E}[\tau_i^r \mid \text{RCT } r] \sim G_\tau \qquad \alpha_r = \mathbb{E}[\alpha_i^r \mid \text{RCT } r] \sim G_\alpha \tag{2}$$

We refer to $(\alpha_r, \tau_r)$ as *structural parameters* because they are population-level counterparts to parameters that describe user-level behavior. We assume that the advertiser, and by extension the ad platform, seeks to estimate each $\tau_r$.

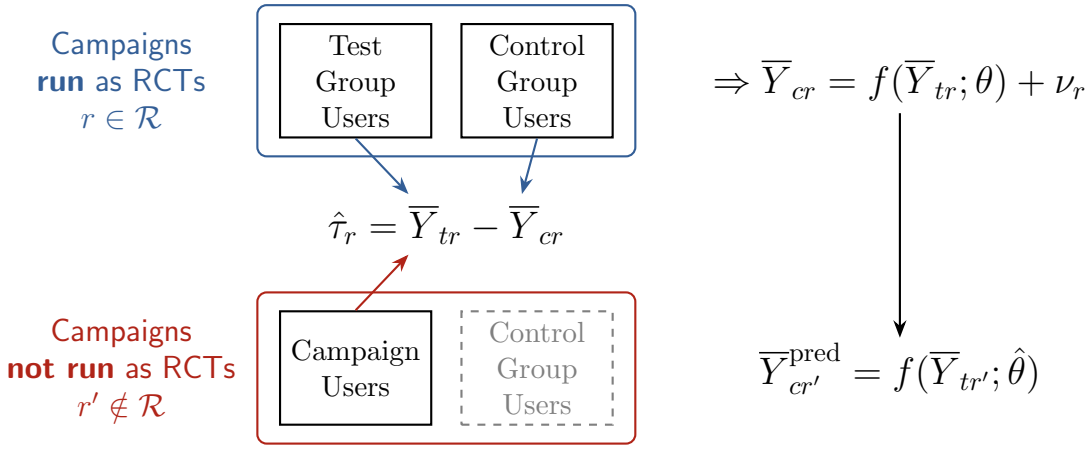
### 3.2 PIE with Perfect Compliance

For simplicity, we begin by assuming perfect compliance in advertising exposure such that all test users are exposed and no control users are exposed. Thus, exposure equals assignment ($D_i^r = Z_i^r$), so the difference-in-means estimator yields an unbiased estimate of the ATE:

$$\hat{\tau}_r = \underbrace{\frac{1}{N_{tr}} \sum_{i \in \text{test}} Y_i^r}_{\overline{Y}_{tr}} - \underbrace{\frac{1}{N_{cr}} \sum_{i \in \text{control}} Y_i^r}_{\overline{Y}_{cr}}. \tag{3}$$

Figure 1 summarizes the problem and our general solution. The problem is that we do not observe $\overline{Y}_{cr'}$ for non-experimental campaigns $r' \notin \mathcal{R}$—that is, counterfactual mean outcomes under nonexposure. At the campaign level, our data are complete for RCTs in that we observe average outcomes for both test and control groups, $\{(\overline{Y}_{cr}, \overline{Y}_{tr})\}_{r=1}^R$. However, for non-experimental campaigns, we have incomplete data, observing only $\overline{Y}_{tr'}$. Our solution is to use the sample of RCTs to train a predictive model, $f(\,\cdot\,; \theta)$, at the campaign level to predict the unobserved $\overline{Y}_{cr'}$.

Figure 1: Overview of the PIE approach.




*Notes:* This diagram illustrates the PIE approach in its simplest form. For campaigns run as RCTs, the causal effect of advertising, $\hat{\tau}_r$, is estimated as the difference between average outcomes in the test and control groups. PIE uses these paired observations across campaigns to learn a mapping, $f(\cdot\,; \theta)$, from test group outcomes to control group outcomes. For campaigns not run as RCTs, the control group is unobserved (dashed box). The trained function $f(\cdot\,; \hat{\theta})$ predicts the counterfactual mean outcome that a control group would have produced, enabling estimation of incremental effects without running an RCT. We refer to right-hand side variables, such as $\overline{Y}_{tr}$, as *post-determined features* because they are only known after the RCT campaign is complete.

The simplest PIE model estimates a regression of this form:

$$\overline{Y}_{cr} = f(\cdot\,; \theta) + \nu_r = \theta_0 + \theta_1 \overline{Y}_{tr} + \nu_r, \tag{4}$$

where $\nu_r$ is assumed to be mean zero. In our model, $\overline{Y}_{tr}$ is an example of a *post-determined feature* because it is only known after the RCT campaign is complete. We refer to $\theta$ as a predictive parameter because it does not have a causal interpretation nor does it directly describe user-level behavior as in the potential outcomes in equation (1). We use a linear regression to clarify the intuition, but the underlying logic can be applied to more sophisticated machine learning models.[7] After fitting the model to obtain the estimate $\hat{\theta}$, we can generate predictions for non-experimental campaigns $r' \notin \mathcal{R}$:

$$\overline{Y}^{\text{pred}}_{cr'} = \hat{\theta}_0 + \hat{\theta}_1 \overline{Y}_{tr'}\ , \tag{5}$$

where the feature $\overline{Y}_{tr'}$ is derived from all the users in the non-experimental campaign, as there is no distinction between a test and control group. This allows us to calculate:

$$\hat{\tau}_{r'} = \overline{Y}_{tr'} - \overline{Y}^{\text{pred}}_{cr'}\ . \tag{6}$$

[7] We use a Random Forest for our empirical application in Section 5.

This approach requires three important assumptions. First, we assume that the test group of an RCT campaign would generate the same observable features if the campaign had instead been run without an RCT, i.e., $\overline{Y}_{tr} \sim \overline{Y}_{tr'}$. This assumption is typically satisfied because the decision to run an RCT affects only whether a control group is withheld from ad delivery. There are no changes in how the platform identifies or serves ads to targeted users. Test group users are unaware they are in an experiment and experience the same ad delivery process that would operate in a regular campaign. Second, the sample of RCT campaigns is representative of the non-experimental campaigns for which we want to predict. This assumption is more likely to hold when the platform controls experiment assignment—for instance, by running experiments in the background without advertiser involvement, thereby avoiding selection based on advertiser characteristics or campaign performance expectations. When instead advertisers self-select into running experiments, representativeness becomes an empirical question that should be assessed before deploying a PIE model (we discuss this in more detail in Section 5). Third, there is no concept shift, meaning that the relationship captured in the PIE model ($f(\cdot;\theta)$) is invariant from the sample used to estimate $\theta$ to the set of non-experimental campaigns used to generate predictions.

To understand the sources of predictive power for the model in equation (4), substitute the expression for user-level outcomes in equation (1) into the group means. By the Law of Large Numbers, as the sample size of each RCT grows ($N_r \to \infty$), each group mean converges to its population expectation plus a sampling error:

$$\overline{Y}_{tr} \approx \mathbb{E}[Y_i^r \mid D_i^r = 1] + \varepsilon_{tr} = \alpha_r + \tau_r + \varepsilon_{tr}, \tag{7}$$

$$\overline{Y}_{cr} \approx \mathbb{E}[Y_i^r \mid D_i^r = 0] + \varepsilon_{cr} = \alpha_r + \varepsilon_{cr}, \tag{8}$$

where $\varepsilon_{tr}$ and $\varepsilon_{cr}$ are independent with $\mathbb{E}[\varepsilon_{tr}] = \mathbb{E}[\varepsilon_{cr}] = 0$. Note that we rely on perfect compliance for $\overline{Y}_{tr}$ because $D_i^r = 1$ for all test users.

Substituting the above into the simple PIE model in equation (4):

$$\underbrace{\alpha_r + \varepsilon_{cr}}_{\overline{Y}_{cr}} = \theta_0 + \theta_1 \underbrace{(\alpha_r + \tau_r + \varepsilon_{tr})}_{\overline{Y}_{tr}} + \nu_r \tag{9}$$

This shows that the model should have some predictive power because $\overline{Y}_{tr}$ and $\overline{Y}_{cr}$ are correlated through the common term $\alpha_r$, the average baseline probability of conversion in RCT $r$. The predictive accuracy of the model depends on how strongly the signal in $\alpha_r$ can be isolated from the other components. Heterogeneity in $\alpha_r$ plays an important role in aiding explanatory power, whereas $\tau_r$ contributes noise. In Section 3.4, we discuss in more detail the conditions under which PIE should perform better or worse.

### 3.3 PIE with Imperfect Compliance and Intermediate Behaviors

The model in the previous section assumed perfect compliance and no intermediate behaviors. We now relax both assumptions by introducing two new variables: (1) an *unobservable* $U_i^r$ that generates one-sided noncompliance in exposure and (2) an *observable* $C_i^r$ that represents intermediate behaviors along a customer's conversion path.

First, users are still randomly assigned to be *eligible* to be exposed through $Z_i^r$, but within the test group, exposure status $D_i^r \in \{0, 1\}$ is a non-random outcome. This outcome now depends on $U_i^r$, which is observed by user $i$ but unobserved by the ad platform. For example, users with higher values of $U_i^r$ may be more engaged with online browsing and thus more likely to encounter ads. These users may also have higher baseline propensities to purchase (higher $\alpha_i^r$). Such correlated unobservables are common in digital advertising (see Lewis et al. 2011) and one of the reasons why many researchers advocate for RCTs to measure such effects (Johnson, 2023).

Formally, we follow Heckman and Vytlacil (2005) and Waisman and Gordon (2025) to include $U_i^r$ in the model by modifying exposure:

$$D_i^r = \mathbb{1}\left\{U_i^r \geq 1 - p_r(Z_i^r)\right\}, \tag{10}$$

where $p_r(Z_i^r)$ governs the process of selection into exposure.[8] Under standard conditions (see Waisman and Gordon 2025), we can normalize $U_i^r \sim \text{Uniform}(0, 1)$. The probability of exposure in the test group is

$$\Pr\left(D_i^r = 1 | Z_i^r = 1\right) = p_r(Z_i^r) \equiv \phi_r \in (0, 1)\ , \tag{11}$$

whereas the exposure probability in the control group is fixed at $p_r(0) = 0$.

Second, advertising platforms typically observe various *intermediate behaviors* of users along their path to an eventual conversion $Y_i^r$. These include searching for a product, visiting a product's detail page, or clicking on an ad. These behaviors may depend on the unobservable $U_i^r$ and may have their own direct effect on $Y_i^r$. For example, a user who clicks on an ad may learn something about the product that increases their likelihood of purchase, independent of the direct effect of ad exposure.

For simplicity, consider a single binary intermediate behavior such that $C_i^r \in \{0, 1\}$, which we interpret as user $i$'s click on an ad for campaign RCT $r$. Because users can only click on an ad if they were exposed to the ad, $C_i^r = 1$ requires that $D_i^r = 1$. The probability of clicking is:

$$\Pr(C_i^r = 1 \mid D_i^r = 1, Z_i^r = 1, U_i^r) = \chi(U_i^r) \tag{12}$$

where $\chi(\cdot)$ captures the notion that clicking may depend on a user's $U_i^r$.

With these new elements in place, we continue to refer to potential outcomes as $Y(D)$ rather

[8] Notice that in this exposure formulation, higher $U_i^r$ implies a greater likelihood of exposure. The more common form for equation (10) is $\mathbb{1}\left\{U_i^r \leq p_r(Z_i^r)\right\}$, which flips the interpretation of $U_i^r$.

than $Y(Z, D, C)$. Using this notation, potential outcomes for unexposed and exposed users are:

$$\begin{aligned} Y_i^r(0) &= \alpha_i^r + \gamma_r U_i^r \\ Y_i^r(1) &= Y_i^r(0) + \tau_i^r + \delta_r C_i^r \end{aligned} \tag{13}$$

where $\tau_i^r$ is the *direct* effect of ad exposure, $\gamma_r$ captures unobserved heterogeneity in baseline outcomes associated with $U_i^r$, $\delta_r$ captures the direct effect of clicking on $Y_i^r$, and $C_i^r \equiv C_i^r\,(D_i^r = 1)$ denotes the potential click status if a user is exposed. Combining these expressions yields observed outcomes:

$$Y_i^r = Y_i^r(D_i^r) = \alpha_i^r + \gamma_r U_i^r + \tau_i^r D_i^r + \delta_r C_i^r \ .$$

We could regress $\overline{Y}_{cr}$ on $\overline{Y}_{tr}$, as we did in equation (4), and use this model to predict $\overline{Y}_{cr'}^{\text{pred}}$. However, the advertiser may be interested in other causal quantities, such as the average treatment effect on the treated (ATT) for campaign $r$:

$$\psi_r \ \equiv \ \text{ATT}_r \ = \ \mathbb{E}[\, Y_i^r(1) - Y_i^r(0) \,|\, D_i^r = 1\,] \,.$$

We can estimate the ATT under standard assumptions for analyzing RCTs with one-sided non-compliance using

$$\hat{\psi}_r \ = \ \frac{\overline{Y}_{tr} - \overline{Y}_{cr}}{\overline{D}_{tr}}, \qquad \overline{D}_{tr} \ = \ \frac{1}{N_r^{\text{t}}} \sum_{i \in \text{t}} D_i^r. \tag{14}$$

In this setting, exposed test-group users coincide with compliers, so equation (14) equals the local average treatment effect for compliers, or the ATT for exposed users (Imbens and Angrist, 1994).

Predictions of $\hat{\psi}_r$ for non-experimental campaigns could be obtained in two steps: first by fitting a PIE model and then by substituting a prediction for $\overline{Y}_{cr}$ into equation (14). However, it is more effective to train a modified version of the original PIE model that directly predicts the causal quantity of interest[9]:

$$\hat{\psi}_r = \theta_0 + \theta_1 \overline{Y}_{tr} + \nu_r \ . \tag{15}$$

Define the direct effect of ad exposure and click rate among the exposed as:

$$\bar{\tau}_r^E \ \equiv \ \mathbb{E}\,[\tau_i^r \,|\, D_i^r = 1]\,, \qquad \bar{C}_r^E \ \equiv \ \mathbb{E}\,[C_i^r \,|\, D_i^r = 1]\,.$$

The ATT (our target) can be decomposed as the sum of the direct effect of exposure and the subsequent effect of clicking among those users who click:

$$\psi_r \equiv \text{ATT}_r = \bar{\tau}_r^E + \delta_r\, \bar{C}_r^E. \tag{16}$$

[9]Predicting $\hat{\psi}_r$ directly aligns the model's objective with the causal estimand of interest. This also avoids error amplification that might inflate variances and add nonlinearity.

With sufficiently large samples of users within each RCT, the group mean outcomes are approximately:

$$\overline{Y}_{tr} \approx \alpha_r + \gamma_r\,\mathbb{E}[U_i^r] \;+\; \underbrace{\phi_r\,\bar{\tau}_r^E}_{\mathbb{E}[\tau_i^r D_i^r|Z=1]} \;+\; \underbrace{\phi_r\,\delta_r\,\bar{C}_r^E}_{\mathbb{E}[\delta_r C_i^r|Z=1]} \;+\; \varepsilon_{tr},$$
$$\overline{Y}_{cr} \approx \alpha_r + \gamma_r\,\mathbb{E}[U_i^r] \;+\; \varepsilon_{cr}.$$

Thus the ATT estimator in equation (14) satisfies:

$$\hat{\psi}_r = \frac{\overline{Y}_{tr} - \overline{Y}_{cr}}{\phi_r} \;\approx\; \bar{\tau}_r^E \;+\; \delta_r\,\bar{C}_r^E \;+\; \frac{\varepsilon_{tr} - \varepsilon_{cr}}{\phi_r}.$$

Substituting into equation (15) yields the population-level relationship:

$$\underbrace{\bar{\tau}_r^E + \delta_r\,\bar{C}_r^E + \frac{\varepsilon_{tr} - \varepsilon_{cr}}{\phi_r}}_{\approx\hat{\psi}_r} = \theta_0 + \theta_1 \underbrace{\left(\alpha_r + \gamma_r\mathbb{E}[U_i^r] + \phi_r\bar{\tau}_r^E + \phi_r\delta_r\bar{C}_r^E + \varepsilon_{tr}\right)}_{\approx\overline{Y}_{tr}} + \nu_r\,. \tag{17}$$

As before, the left- and right-hand sides share common structural components. Changing the target of the prediction model from $\overline{Y}_{cr}$ to $\hat{\psi}_r$ simply alters the structural parameters that determine the sources of predictive power. In equation (9) of the full compliance model, heterogeneity in $\alpha_r$ played a critical role in aiding explanatory power, while heterogeneity in $\tau_r$ contributed noise. Here, the situation is reversed, with predictive signal driven by cross-campaign variation in $(\phi_r, \bar{\tau}_r^E, \delta_r, \bar{C}_r^E)$, while dispersion in $(\alpha_r, \gamma_r)$ only enters through $\overline{Y}_{tr}$ and thus primarily contributes noise.

In practice, an advertising platform observes many campaign-level signals beyond $\overline{Y}_{tr}$, such as exposure rate, click-through rate, or last-click conversions, all of which may encode information about the structural parameters that determine treatment effects. The next section uses simulations to characterize when these features improve prediction.

### 3.4 When should we expect PIE to work better or worse?

In this section we identify the conditions under which PIE performs better or worse in predicting incremental effects $\hat{\psi}_r$. Above, we have already provided some intuition for the variation in structural parameters that allows PIE to be predictive of treatment effects. To explore how a more general implementation of the PIE model performs, we simulate a model of full compliance (as in Section 3.2), followed by a model of partial compliance that allows for unobserved heterogeneity to be correlated with selection into treatment and conversions (as in Section 3.3). We summarize when one should expect PIE to perform better or worse in predicting incremental effects. Appendix 3 presents a more detailed version of this section.

### 3.4.1 Overview of Simulation Setup

Each simulation constructs the potential outcomes in equation (13) for a set of campaigns $r \in \mathcal{R}$, each containing $N_r$ users. We specify the distribution across campaigns of $\alpha_r$ (baseline outcomes), $\gamma_r$ (effect of user unobservable on baseline outcomes), $\tau_r$ (treatment effect of exposure), $\text{corr}(\alpha_r, \tau_r)$, $\delta_r$ (treatment effect of clicking), the probability of exposure $\phi_r$, and users' base-level probability of clicking $\Pr(C_r)$. We use these to simulate user-level behaviors for each campaign. This yields for each campaign $r$ a user-level ad-exposure status $D_i^r \in \{0,1\}$, an ad-click $C_i^r \in \{0,1\}$, and control and test group outcomes $Y_{ci}^r \in \{0,1\}$ and $Y_{ti}^r \in \{0,1\}$. In addition, we define a last-click conversion as $LCC_i^r = Y_{ti}^r \cdot C_i^r$, indicating that user $i$ both clicked and converted in the test condition.[10]

We aggregate individual-level data at the campaign level, yielding for each campaign $r$ the number of conversions in each group $(Y_{tr}, Y_{cr})$, clicks $(C_{tr} \geq 0,\ C_{cr} = 0)$, last-click conversions $(LCC_{tr} \geq 0,\ LCC_{cr} = 0)$, and the ad-exposure rate $(\phi_{tr} \geq 0,\ \phi_{cr} = 0)$.

We divide campaigns into equal-sized training and test samples. Using the training sample, we estimate a series of PIE specifications via OLS regression, progressively adding covariates until we arrive at the following full specification:

$$\hat{\psi}_r = \theta_0 + \theta_1 Y_{tr} + \theta_2 \phi_{tr} + \theta_3 C_{tr} + \theta_4 LCC_{tr} + \nu_r \ . \tag{18}$$

Note that the right-hand side features rely only on data from the test group. After obtaining $\hat{\theta}$ from the training sample, we generate predictions $\psi_r^{\text{PIE}}$ by applying the fitted model to campaigns in the test sample.

We evaluate prediction accuracy using out-of-sample $R^2$ between $\hat{\psi}_r$ and $\psi_r^{\text{PIE}}$. Higher $R^2$ indicates that the model substantially outperforms a naive baseline that predicts the mean treatment effect, $\bar{\psi}$, for every campaign; $R^2 = 0$ indicates no improvement over predicting the mean. We report $R^2$ for models with increasingly rich feature sets: total conversions only $(R^2_{Y_t})$, adding exposure fraction $(R^2_{Y_t,\phi})$, adding clicks $(R^2_{Y_t,\phi,C})$, and the full model including last-click conversions $(R^2_{Y_t,\phi,C,LCC})$. Each simulation is repeated 100 times with different random seeds, and we pool all test observations across repetitions to calculate a single $R^2$ for each scenario.

### 3.4.2 Simulation Results for Full Compliance

We begin with the simple case described in Section 3.2, where we assumed perfect compliance with advertising exposure and no intermediate behaviors. To obtain the equivalent of equation (9) but directly predicting $\hat{\psi}_r$ instead of the control group outcome $\overline{Y}_{cr}$, we simplify equation (17) assuming perfect compliance ($\phi_r = 1$), no effect of an unobservable $U_i^r$ ($\gamma_r = 0$), and no effect of intermediate behaviors $C_i^r$ ($\delta_r = 0$). This yields

$$\underbrace{\tau_r + \varepsilon_{tr} - \varepsilon_{cr}}_{\hat{\psi}_r = \hat{\tau}_r = \overline{Y}_{tr} - \overline{Y}_{cr}} = \theta_0 \ + \ \theta_1 \underbrace{\left(\alpha_r + \tau_r + \varepsilon_{tr}\right)}_{\overline{Y}_{tr}} + \nu_r \tag{19}$$

[10] Since the simulation does not have a time dimension, there is no attribution window $w$ associated with a last-click conversion.

This formulation suggests that heterogeneity in $\tau_r$ should play an important role in aiding explanatory power, while heterogeneity in $\alpha_r$ should contribute noise.

To simulate these effects, we specify a data-generating process with perfect compliance ($p_E = 1$), heterogeneous baseline conversions across campaigns, heterogeneous treatment effects, and no effect of clicks on conversions ($\delta_r = 0$). See Appendix 3 for the exact simulation parameters.

First, we vary the number of RCTs across four values (50, 100, 1,000, and 10,000), holding constant the sample size of each RCT at 10,000 users. Next, we fix the number of RCTs at 1,000 while varying the size of each RCT (1000, 5,000, 10,000 users). We find that PIE benefits from increased sample size: both more RCTs and larger RCTs improve performance by reducing estimation noise (Table A-3 on page A-6). For the remainder of the simulations, we use 1,000 RCTs with 10,000 users each.

Second, we vary the heterogeneity in baseline conversion probability $\alpha_r$ and in the treatment effect $\tau_r$. Consistent with the theoretical prediction above, greater variation in baseline conversion rates $\alpha_r$ harms performance because it adds noise to the predictors without providing information about $\tau_r$. We also find that greater variation in treatment effects $\tau_r$ improves $R^2$ because it creates more opportunity to outperform the naive baseline of predicting the mean (Table A-4 on page A-6).

Third, we analyze how the correlation between $\alpha_r$ and $\tau_r$ affects PIE's predictive accuracy. Recall from the right side of equation (19) that observed treatment group conversions depend on both $\alpha_r$ and $\tau_r$. We find that a positive correlation amplifies the signal in $Y_{tr}$ (Table A-5 on page A-7). A strong negative correlation means that in $Y_{tr} = \alpha_r + \tau_r + \varepsilon_{tr}$, $\alpha_r$ and $\tau_r$ nearly cancel each other out, leaving only noise in $Y_{tr}$ and eliminating PIE's predictive power.

After this exploration, we revert to $\text{corr}(\alpha_r, \tau_r) = 0$ but now allow clicks to affect conversions ($\delta_r > 0$). We find that under low click probability, introducing $\delta_r$ heterogeneity has minimal impact. When few users click, these features contain little information regardless of click effect variation (Panel A in Table A-6 on page A-8). Under high click probability, PIE performs better relative to a simple mean prediction as $\delta_r$ heterogeneity increases. The incremental contribution of click-based features grows substantially with $\delta$ heterogeneity (Panel B in Table A-6 on page A-8).

### 3.4.3 Simulation Results for Partial Compliance

We now simulate a data-generating process under partial compliance in exposure. Recall from equation (10) that exposure is determined by $D_i^r = \mathbb{1}\{U_i^r \geq 1 - p_r(Z_i^r)\}$, where $U_i^r \sim \text{Uniform}(0, 1)$. The parameter $\gamma_r$ controls how strongly $U_i^r$ affects baseline conversion probability. When $\gamma_r > 0$, the same unobservable that determines exposure also affects baseline conversion rates, leading to exposed consumers who have higher baseline conversion probabilities than unexposed consumers. This is an empirical identification challenge for observational models.

First, we vary the heterogeneity of $\gamma_r$. We find that PIE performs less well with high heterogeneity in $\gamma_r$. This is analogous to increasing heterogeneity in the baseline probability $\alpha_r$: greater heterogeneity in $\gamma_r$ increases the variance of $Y_{tr}$ without proportionally increasing our ability to predict $\tau_r$ because $\gamma_r$ appears only on the right-hand side of equation (17) (Panel A in Table A-7

on page A-9).

Second, we vary the intensity of selection into exposure. We find that PIE's post-determined features beyond total conversions—exposure fraction, clicks, and last-click conversions—become increasingly valuable in predicting treatment effects as exposure becomes more selected (lower $p_E$). Not only does the $R^2$ of the full PIE model increase, indicating that PIE outperforms a simple prediction of the mean treatment effect, but the full model also improves relative to a simple PIE model that uses only $Y_{tr}$ for prediction (Panel B in Table A-7 on page A-9). This improvement occurs because, as selection increases, only users with increasingly high $U_i^r$ are exposed, and post-determined features become increasingly informative about user composition, which PIE can exploit relative to predicting the mean.

Third, we find that click-based features improve predictions even when clicks have no direct causal effect on conversions ($\delta_r = 0$) (Panel C in Table A-7 on page A-9). This occurs because, under partial compliance, clicks reveal information about user composition. Users with high $U_i^r$ are both more likely to be exposed and, through the $\gamma_r$ parameter, more likely to convert. Since click probability also depends on $U_i^r$ (equation 12), click features serve as a proxy for the unobserved user characteristics that drive both exposure and outcomes. Intuitively, even if $\delta_r = 0$, the click event $C_i^r$ lies on a causal path from $U_i^r$ (through exposure $D_i^r$) and is therefore informative about the latent factors that jointly determine exposure and baseline conversion propensity. Conditioning on click features thus improves prediction of treatment effects by revealing information about user composition, not by capturing any direct click-to-conversion channel.

### 3.4.4 Simulation Summary

In summary, PIE performs best relative to predicting the mean treatment effect when: (1) treatment effects vary substantially across campaigns, (2) baseline outcomes vary modestly, (3) baseline outcomes and treatment effects are positively correlated, (4) the number and sample sizes of RCTs are large, (5) there is selection into exposure, and (6) the correlation between the selection mechanism and outcomes is similar across campaigns. PIE performs worst relative to predicting the mean treatment effect when treatment effects are homogeneous, baseline outcomes vary widely, or baseline outcomes and treatment effects are negatively correlated.

Having characterized when PIE should perform better or worse, translating these findings into a specific real-world setting is challenging absent detailed data from RCTs. Given access to such data, we can evaluate some of the constructs that affect PIE's performance. However, once such data are available, we recommend directly evaluating PIE using out-of-sample performance metrics, as we do in the following sections.

## 3.5 General PIE Model

The models in Sections 3.2 and 3.3, and the simulations in Section 3.4, served a specific purpose. They demonstrated, within a stylized potential outcomes framework, that campaign-level features can predict treatment effects because both the features and the causal quantities are functions of

common structural parameters. The true model of user behavior is almost certainly more complex than any tractable formalization—involving additional unobservables, numerous intermediate behaviors, and interactions among them. One benefit of the PIE model is that the researcher need not correctly specify this data-generating process. Because the RCTs have already identified the causal effects, the prediction step inherits no identification burden. The researcher can include as features any campaign-level variables that the platform records for both experimental and non-experimental campaigns. The supervised learning model determines which features carry predictive information.

This motivates a general formulation. We introduce an aggregated feature vector $\mathbf{X}_r = \big(\mathbf{X}_r^{\text{pre}},\ \mathbf{X}_r^{\text{post}}\big)$ partitioned into pre-determined variables, fixed *ex-ante* before the campaign runs, and post-determined variables, realized *ex-post* when the campaign concludes. The model becomes an arbitrary, possibly nonlinear function $f(\mathbf{X}_r;\theta)$. Let $\hat{\psi}_r$ denote the campaign-level causal target (e.g., ATT or a transformation). We write:

$$\hat{\psi}_r = f\big(\mathbf{X}_r;\theta\big) + \nu_r, \qquad \mathbb{E}[\nu_r \mid \mathbf{X}_r] = 0. \tag{20}$$

As a special case, the one-regressor PIE model in Section 3.2 had $\mathbf{X}_r = \big(\mathbf{X}_r^{\text{pre}},\ \mathbf{X}_r^{\text{post}}\big) = \big(\emptyset,\ \overline{Y}_{tr}\big)$. The mean outcome in the test group, $\overline{Y}_{tr}$, is an example of a post-determined feature because it is known only after the campaign concludes and is formed by aggregating user-level variables.

More generally, post-determined features such as exposure rate, click-through rate, or last-click conversions are informative because they too depend on the behavioral and structural factors that drive treatment effects.[11] These metrics depend on actual conversion outcomes ($Y_i^r$) as in the case of last-click conversions, or on intermediate behaviors ($C_i^r$) as in the case of click-through rates, or both. We can express these features generically using:

$$\mathbf{X}_r^{\text{post}} = \frac{1}{N_r}\sum_{i=1}^{N_r} h\big(Y_i^r,\ D_i^r,\ C_i^r,\ \text{aux}_i^r\big) \tag{21}$$

where $h(\cdot)$ is a vector-valued mapping and $\text{aux}_i^r$ denotes any additional user-level signals (e.g., page views, ad placement indicators, invalid clicks) that the platform logs.

In addition, pre-determined features $\mathbf{X}_r^{\text{pre}}$ that include campaign configuration options (e.g., objective, creative, budget, targeting options) or advertiser-level descriptors (e.g., vertical) may shift the distribution of unobservables or the shape of treatment heterogeneity, so including them in $f(\cdot)$ helps the model adapt to concept shift across advertisers. The platform need only ensure that it logs the necessary metrics from all campaigns with which to train the model.

[11] Because PIE is a prediction problem, the temporal ordering of features and the target variable is irrelevant for validity. Lower-funnel outcomes (e.g., purchases) are valid predictors of upper-funnel treatment effects, and vice versa, because the model exploits correlations across campaign-level aggregates rather than within-user causal pathways.

# 4 Application to Experiments at Meta

Having characterized the theoretical conditions under which PIE should perform well or poorly, we now turn to implementation and evaluation. In this section, we describe the dataset of Meta ad experiments to which we apply the general version of the PIE model described in Section 3.5 and explain our specific estimation and evaluation approach. Results follow in Section 5.

## 4.1 Data

The data consist of ad experiments run by advertisers using Meta's "Conversion Lift" product, which delivers ads across Facebook, Instagram, and the Facebook Audience Network.[12] Each experiment randomly assigns users in the advertiser's target audience to test (eligible for exposure) or control (ineligible) groups, after which delivery occurs via standard auctions.[13] The experiments in our data are a superset of those in Gordon et al. (2023).[14]

We randomly selected experiments representative of large-scale tests run by U.S. advertisers between November 1, 2019, and March 1, 2020, with at least one million test group users and at least 5,000 test group conversions.[15] The resulting sample contains 839 experiments across different industry verticals with varying targeting choices, campaign objectives, conversion outcomes, and budgets. An advertiser can run multiple treatment-control pairs in a multi-cell test to evaluate different strategies. We treat each cell (treatment-control pair) as a unique experiment since populations do not overlap by construction. In our sample, 120 experiments include multiple pairs, yielding 998 unique treatment-control pairs, which we refer to as experiments going forward.

The data contain approximately 11.4 billion user-experiment observations with 52.5 billion ad impressions. The median experiment runs for 28 days, includes 7.4 million users, and delivers 21.5 million impressions. The median control group share is 10%, and the median test group exposure rate is 76.3%. Our sample spans many industry verticals, with the e-commerce and retail verticals comprising roughly half of all RCTs. Distributional summaries of duration, reach, splits, exposure rate, and impressions are shown in Appendix Figure A-1.

Most experiments measure multiple conversion outcomes (e.g., purchases, checkout initiations, app installs, page views). While advertisers choose which events to measure, these typically fall into three funnel stages: upper funnel (e.g., page views), mid-funnel (e.g., cart additions), and lower funnel (e.g., purchases). Our 998 experiments measure 2,226 conversion events across the funnel. We refer to each experiment-conversion event pair as an RCT and classify RCTs as Lower Funnel (773), Mid Funnel (713), or Upper Funnel (740).

We estimate the average treatment effect on the treated (ATT) for each RCT using the Wald/2SLS estimator with randomized assignment as the instrument (Imbens and Angrist, 1994), which un-

[12] `https://www.facebook.com/business/help/221353413010930?id=547299432790676`, accessed on October 14, 2025.

[13] For details on ad delivery and advertising RCTs at Meta, see Section 2 of Gordon et al. (2023).

[14] Data requirements differed between these papers, allowing us to include experiments that were excluded from Gordon et al. (2023).

[15] These thresholds help balance statistical power and computational feasibility.

der one-sided noncompliance and monotonicity identifies the average effect for exposed users. We compute the implied number of (estimated) incremental conversions as

$$\mathrm{IC}_r = \widehat{\mathrm{ATT}}_r \cdot \overline{D}_{tr} \cdot N_{tr} \ , \tag{22}$$

that is, estimated ATT times the number of exposed users in the test group.

Consistent with Section 3.5, the features available for non-experimental campaigns must be computable without data from the control group. For post-determined features entering $\mathbf{X}_r^{\mathrm{post}}$, we use (1) the mean outcome in the test group $\overline{Y}_{tr}$; (2) the exposure rate $\overline{D}_{tr}$; and (3) the number of last-click conversions $\mathrm{LCC}_{tr}(w)$ measured over standard attribution windows $w \in \{1 \text{ hour}, 1 \text{ day}, 7 \text{ day}, 28 \text{ day}\}$.[16] Because the prediction target is expressed in conversions per dollar (ICPD), we normalize $\overline{Y}_{tr}$ and each $\mathrm{LCC}_{tr}(w)$ by total campaign spend so that features and outcome are on a common scale.[17] We note that our data contain fewer post-determined features than a platform would typically observe.

Finally, for pre-determined features that enter $\mathbf{X}_r^{\mathrm{pre}}$, we observe advertiser- and campaign-level features for each experiment, including (1) targeting settings, (2) bidding strategy, (3) optimization settings, (4) advertiser characteristics and platform experience, and (5) actual campaign spending.

Next, we explain the specific implementation and evaluation of the model using these data.

## 4.2 Model Estimation and Evaluation

For the outcome $\hat{\psi}_r$ of the prediction model, we use a simple transformation of the ATT, namely the number of incremental conversions per dollar (ICPD):

$$\mathrm{ICPD}_r = \frac{\mathrm{IC}_r}{\mathrm{Cost}_r} \ . \tag{23}$$

The ICPD is a common financial metric that advertisers use to assess campaign performance and helps normalize treatment effects against different budget levels.[18]

We implement $f\big(\mathbf{X}_r;\theta\big)$ as a random forest and train on various combinations of the pre- and post-determined features enumerated in Section 4.1.[19] We perform hyperparameter tuning for each model specification using 10-fold cross-validation.

[16] Our data do not contain raw clicks $C_i^r$ or total clicks. However, one could interpret the $\mathrm{LCC}_r(w)$ as being formed from the interaction of a conversion event $Y_i^r$ and a click event $C_i^r$, constrained to occur within a window $w$ of each other.

[17] The exposure rate $\overline{D}_{tr}$ is already a proportion and does not require this normalization.

[18] A related metric is the cost per incremental conversion (CPIC), the inverse of ICPD, which advertisers often use to assess profitability. Because advertisers typically have private information on the profit per conversion, CPIC helps them determine whether a campaign produced a positive return on ad spend (ROAS). From a modeling standpoint, however, CPIC is problematic because (i) it is undefined when estimated lift is zero or negative and (ii) it produces extremely skewed error distributions when lift is small. ICPD avoids both issues.

[19] We explored other models. We found that linear regression generally underperformed and that XGBoost performed similarly or slightly better than random forests but was more computationally burdensome.

To train and evaluate the model, we use weighted out-of-sample $R^2$:

$$R^2 = 1 - \frac{\sum_r \omega_r \cdot (\text{ICPD}_r - \text{ICPD}_r^{\text{PIE}})^2}{\sum_r \omega_r \cdot (\text{ICPD}_r - \overline{\text{ICPD}})^2} \tag{24}$$

where $\text{ICPD}_r^{\text{PIE}}$ is the prediction, $\text{ICPD}_r$ is the observed value, $\overline{\text{ICPD}}$ is the mean observed value, and $\omega_r$ is the weight associated with RCT $r$, which sum to one. We use the campaign cost as the weight to reflect the notion that the ad platform prefers higher accuracy on more expensive campaigns.[20] We use a nonparametric bootstrap with 1,000 draws to estimate the distribution of out-of-sample $R^2$.[21]

The preferred evaluation metric and weights are design choices for the advertising platform, and we are not wedded to the particular choices made in this empirical application. The platform could train different models using different evaluation metrics and weights, depending on their use case.

One concern with our evaluation is that the post-determined features $\overline{Y}_{tr}$ and $\overline{D}_{tr}$ also appear in the construction of the dependent variable $\text{ICPD}_r$. This shared dependence on finite-sample realizations can induce mechanical correlation between features and outcomes, potentially inflating out-of-sample $R^2$. The standard remedy is within-campaign sample splitting: using one random half of users to compute the outcome and the other half to compute features, as we implement in our simulations (Appendix 3). Our empirical application uses pre-aggregated campaign-level data, which precludes this approach. However, the large median sample size of our RCTs (7.4 million users) implies that sampling noise in both features and outcomes is small, limiting the scope for mechanical correlation to materially affect our results.

# 5 Results

This section presents baseline results and robustness evaluations from applying our PIE model to the 2,226 RCTs. We first present the model's performance across all RCTs and analyze its predictive accuracy across several campaign characteristics: purchase funnel level, prospecting versus retargeting campaigns, and industry vertical. We also examine performance as a function of the number of RCTs in the training sample.

Second, in Section 5.2, we assess the performance of PIE for new campaigns of advertisers whose prior campaigns were used to train PIE. We compare this to the performance of PIE for new campaigns of advertisers without prior campaigns in PIE's training data. This examines the model's ability to transfer information across advertisers, testing how well the model can address

[20] We have considered weights based directly on the inverse of the ATT's sample variance, including versions based on calculations using methods from the meta-analysis literature to better account for within- and between-experiment variance. We do not find that this materially changes our results.

[21] Because $\text{ICPD}_r$ is estimated, the observed $R^2$ has a ceiling below one: even a perfect model would leave residual variance from estimation noise in the outcomes. Given our sample sizes, this noise may be small relative to the cross-campaign variation in true effects, suggesting a ceiling close to one, though we cannot verify this without decomposing the residual into model error and outcome noise. We thank Randall Lewis for raising this point.

the "cold-start" problem of generalizing to completely new advertisers.

Third, in Section 5.3, we examine the model's ability to transfer information across campaign segments. For example, we assess PIE's ability to extrapolate by predicting campaign performance for large advertisers when the model is trained only on RCTs from smaller advertisers. We compare this predictive performance to the within-segment performance when the model is trained on RCTs from large advertisers.

### 5.1 Baseline

Figure 2 presents the baseline performance of several PIE model specifications. PIE(Pre) includes campaign characteristics (which are pre-determined variables) in $\mathbf{X}_r^{\text{pre}}$ and omits any post-determined variables ($\mathbf{X}_r^{\text{post}} = \emptyset$). Next, in PIE(Pre, $Y_t$) we add one post-determined variable, such that $\mathbf{X}_r^{\text{post}} = \{\overline{Y}_{tr}\}$, while we retain $\mathbf{X}_r^{\text{pre}}$. PIE(Pre, $Y_t$, LCC-7D) adds $\text{LCC}_r$(7 day), the last-click 7-day conversion counts, and PIE(Full) adds the remaining post-determined features. The error bars on the graph represent one standard deviation from the mean $R^2$.

We compare the performance of these specifications to each other and against a benchmark of Raw LCC-7D. This benchmark involves no model: it divides the raw last-click 7-day conversion count by campaign spend and treats the result as if it were ICPD. The resulting $R^2$ measures how well this unadjusted metric tracks true incremental effects. We focus on this attribution window because it is a common choice for advertisers.

We draw several observations from Figure 2. First, adding the first post-determined feature, $\overline{Y}_{tr}$, to the model increases out-of-sample $R^2$ from 0.35 to 0.46 despite the model already conditioning on pre-determined features that describe a campaign's audience targeting, budget, bidding strategy, and advertiser type. Second, including LCC-7D as another post-determined feature leads to a substantial $R^2$ increase to 0.85. Adding more post-determined features in the PIE(Full) specification, such as other LCC metrics, yields a modest increase to $R^2 = 0.88$. The additional LCC metrics likely produce relatively small gains because they are mechanically correlated with each other. More broadly, the limited marginal gains from adding other LCC windows reflect the narrow scope of the available post-determined variables in our dataset. Our data contain fewer post-determined features than a platform would typically observe. A platform designing its data collection to support PIE could incorporate a wider range of behavioral signals (e.g., view-through conversions, engagement metrics, dwell times) that might yield further improvements in predictive accuracy.

Critically, all PIE models outperform the raw LCC 7-day metric. Using the best model, an advertiser increases the $R^2$ of the prediction from 0.19 to 0.88, a substantial increase in predictive performance for treatment effects in campaigns that were not run as RCTs. The PIE predictions also exhibit lower variance compared to using last-click conversions alone. In general, out-of-sample $R^2$ values can be negative; this occurs when the predictor performs worse than simply using the mean outcome. This is particularly true for raw last-click metrics, which involve no model and simply treat the attribution count as if it directly measured incrementality.

The contrast between the strong contribution of LCC-7D within PIE and the weak standalone performance of Raw LCC-7D reflects a fundamental distinction. PIE learns across campaigns how last-click conversions relate to true incremental effects. Raw LCC-7D involves no such learning and simply assumes that attributed conversions equal incremental conversions. Across the full sample, mean LCC-7D per dollar overstates mean ICPD by a factor of 1.33. An OLS regression of $\hat{\psi}_r$ on LCC-7D per dollar yields a slope of 0.69, indicating that on average approximately 69% of LCC-7D per dollar are incremental. Thus, using LCC-7D per dollar as a direct substitute for ICPD introduces a systematic upward bias that is reflected in the out-of-sample $R^2$. The LCC-7D metric nonetheless remains highly correlated with true incremental effects (Spearman $\rho = 0.89$), indicating that it carries substantial information about incrementality. PIE exploits this information by learning the mapping from confounded attribution metrics to causal quantities.

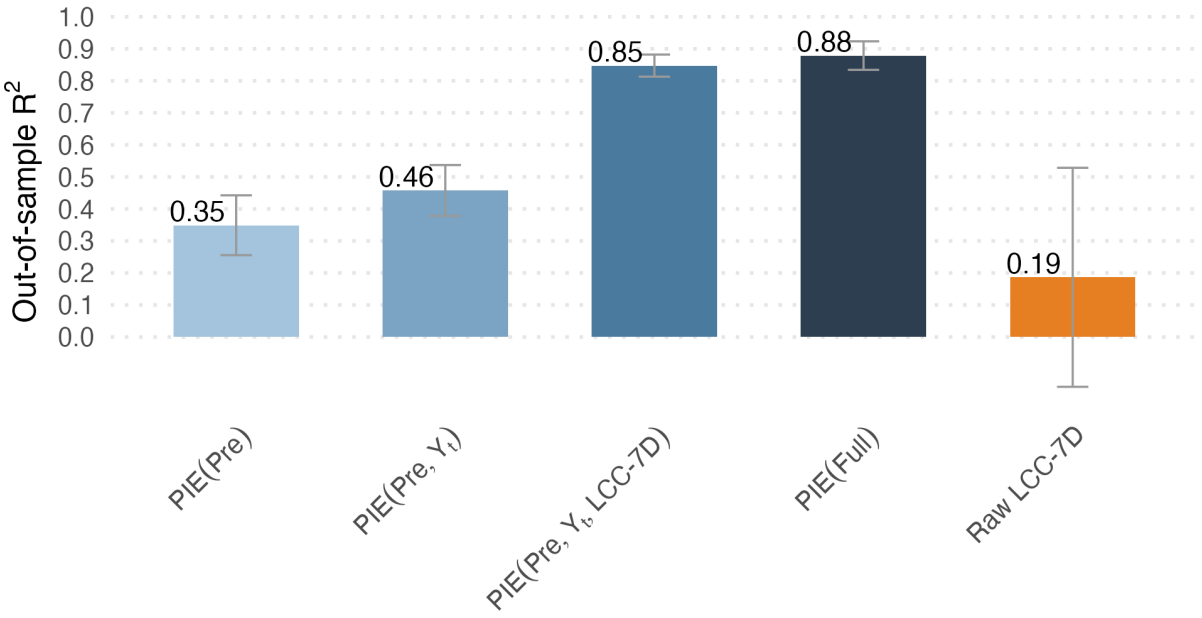
Comparing PIE(Pre, $Y_t$) and PIE(Pre, $Y_t$, LCC-7D) reveals the importance of including last-click 7-day conversions as a predictive feature. The predictive contribution of a last-click metric—often viewed as a biased measure of causal ad effects—likely reflects the nature of ad exposure on social media platforms where users typically visit the platform to browse content, not to search for or purchase specific products. As a result, the endogeneity of the click event is substantially attenuated relative to search advertising. If a user stops scrolling, clicks on an ad, and then purchases within seven days, the click-to-conversion sequence is likely informative about a genuine causal effect of the ad. In the framework of our model, this informativeness arises through two channels. First, the click event $C_i^r$ may have a direct causal effect on the outcome ($\delta_r > 0$), for instance because clicking exposes the user to additional product information. Second, absent a direct effect ($\delta_r = 0$), the click reveals information about the user's unobservable $U_i^r$, which in turn correlates with both exposure and baseline conversion propensity, as demonstrated in our simulations (Panel C of Table A-7).

This differs from search advertising, where users arrive with pre-existing purchase intent. A user who searches for a product, clicks on the first sponsored result, and then purchases generates a last-click conversion that is heavily confounded by the intent that drove the search. In such settings, last-click conversions may be a substantially noisier signal of incremental ad effects. This distinction implies that the strong predictive power of LCC-7D documented here should not be assumed to generalize to search advertising platforms without empirical validation.

Figure 3 decomposes the earlier results by the level of purchase funnel of the conversion event. Across the funnel levels, we observe a pattern consistent with the baseline results: conditioning on $\overline{Y}_{tr}$ improves over using only the pre-determined variables $\mathbf{X}_r^{\text{pre}}$, conditioning on $\text{LCC}_r(7\text{ day})$ yields even larger improvements, and adding the remaining post-determined variables produces modest gains. In this case, using raw last-click conversions (per dollar) yields negative $R^2$, indicating that a prediction based on the mean ICPD across RCTs would be more accurate than using raw last-click conversions to predict ICPD for new campaigns. The full model, in contrast, achieves $R^2$ values of 0.91, 0.92, and 0.85, for lower, mid-, and upper-funnel outcomes, respectively.

Figure 4 separates the results according to the type of audience targeting used in each campaign. Campaigns were classified as "prospecting" if at least 70% of the audience constituted new potential customers for the advertiser. Otherwise, we classified them as "retargeting" campaigns. The results

Figure 2: PIE model performance.




*Notes:* This figure compares the out-of-sample $R^2$ to predict ICPD ($\hat{\psi}_r$). The bars represent progressive additions of features to the PIE model, from PIE(Pre) (using only pre-determined features) to PIE(Full) (using all features). Raw LCC-7D serves as a benchmark that divides the raw last-click 7-day conversion count by campaign spend and treats the result as if it were ICPD, with no model applied. The $R^2$ value is shown above each bar, and error bars represent one standard deviation from the mean $R^2$, estimated via bootstrap.

demonstrate a similar pattern of predictive performance across the campaign types as observed in the earlier figures.[22]

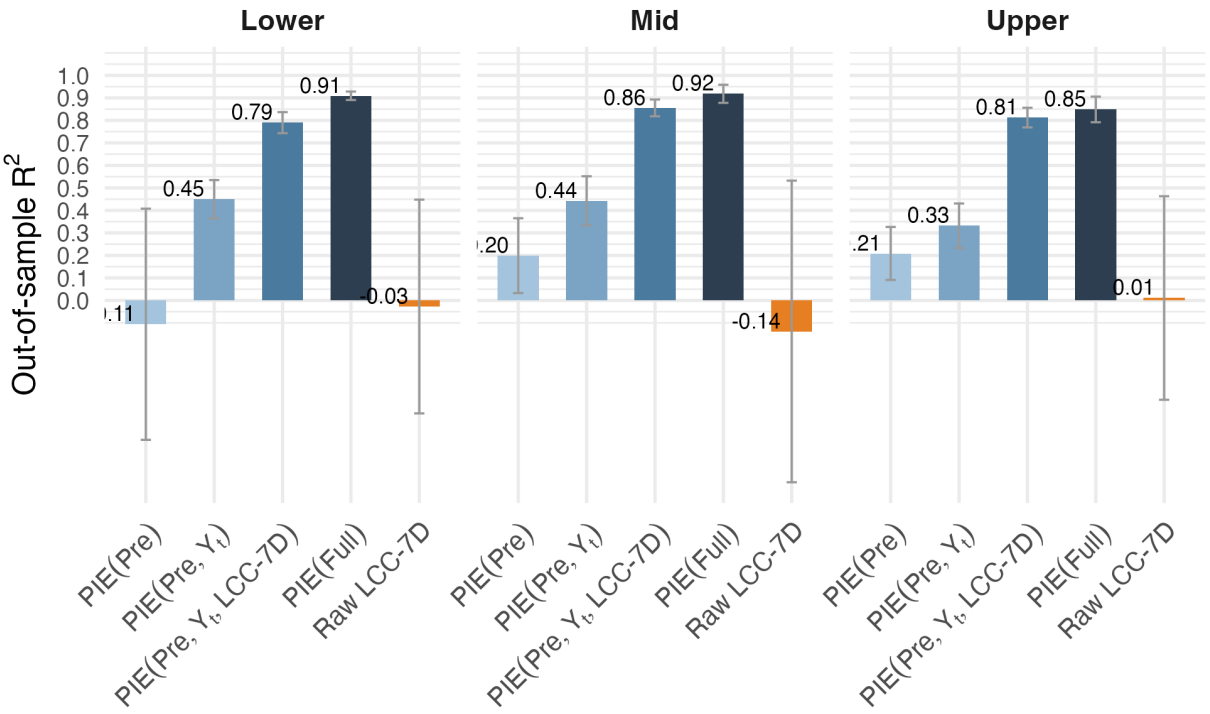
Figure 5 presents the results organized by industry vertical of the advertiser for verticals with at least 200 RCTs. The figure shows that the PIE(Full) model outperforms the raw last-click metric, with the margin varying substantially across verticals. The ratio of mean LCC-7D per dollar to mean ICPD is 1.2 in e-commerce, 1.5 in retail, and 2.0 in travel, and OLS slopes of $\hat{\psi}_r$ on LCC-7D per dollar are 0.76, 0.63, and 0.48, respectively. In travel, approximately half of last-click conversions per dollar would have occurred without the ad. In e-commerce, where the bias is smallest, the raw metric still achieves reasonable out-of-sample performance ($R^2 = 0.90$), whereas in retail and travel, the systematic overstatement produces predictions worse than the sample mean.

Figure 6 investigates the trade-off between the quantity of RCTs and predictive performance. By repeatedly training the PIE(Full) model on random subsamples of increasing size and evaluating its $R^2$, we assess the model's performance with smaller datasets. We observe that, with a dataset of 800 RCTs (just over one-third the size of the full data), the $R^2$ is close to 80%. However, the $R^2$ drops quickly as the sample size shrinks below 400 RCTs.

In a real-world implementation, the platform must determine how to obtain a representative

[22]This classification serves as a rough approximation to infer the campaign type, as it is possible for an advertiser to target a mix of users with any given campaign. The qualitative findings are not overly sensitive to the 70% threshold.

Figure 3: PIE model performance by Conversion Event Funnel




*Notes:* This figure compares the out-of-sample $R^2$ to predict ICPD ($\hat{\psi}_r$), decomposed by the type of outcome according to the conversion event funnel. Lower funnel events mostly represent online purchases, whereas upper funnel conversions are most frequently page landings. The bars represent progressive additions of features to the PIE model, from PIE(Pre) (using only pre-determined features) to PIE(Full) (using all features). Raw LCC-7D serves as a benchmark that divides the raw last-click 7-day conversion count by campaign spend and treats the result as if it were ICPD, with no model applied. The $R^2$ value is shown above each bar, and error bars represent one standard deviation from the mean $R^2$, estimated via bootstrap.

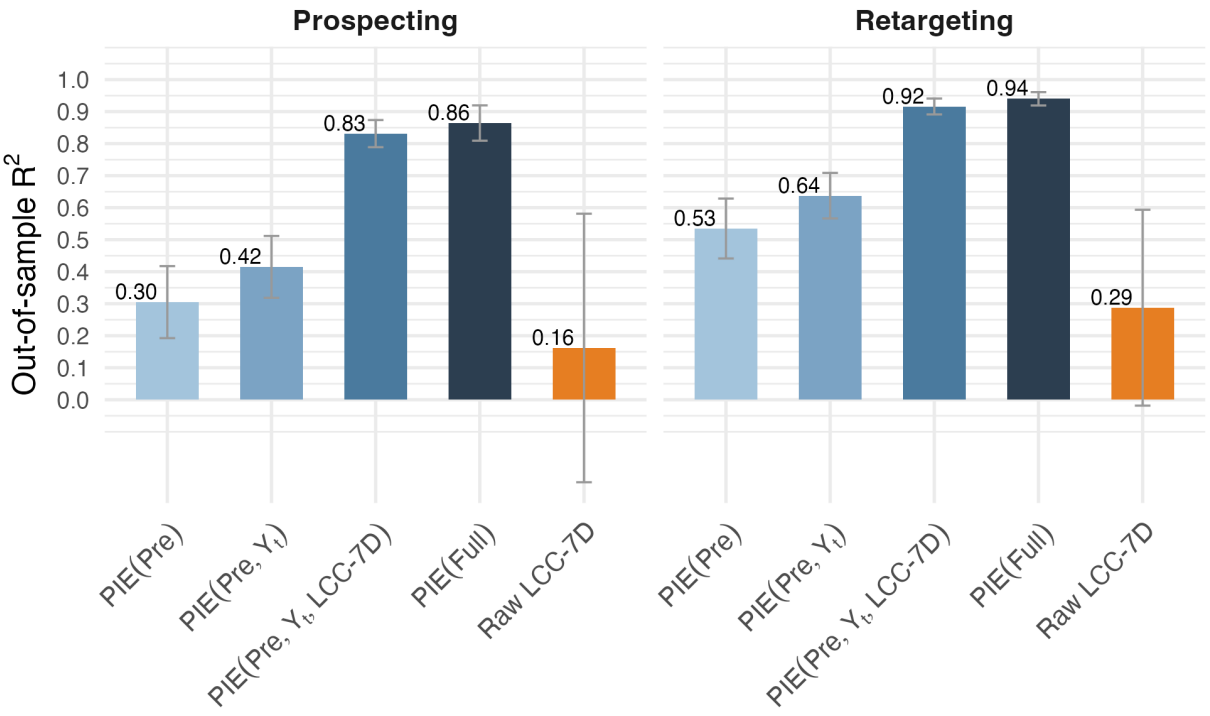
sample of RCTs while balancing the costs of running these experiments (e.g., the opportunity cost of withholding impressions from a control group). Furthermore, the platform must continually assess whether its current training sample has become stale, requiring new experiments to maintain model performance. We discuss these and related issues in Section 7.

### 5.2 Predictive Accuracy for New vs. Existing Advertisers

One requirement for the practical application of the PIE model is its ability to provide accurate predictions for both existing and new advertisers. To test these two dimensions of performance, we implement two complementary cross-validation frameworks.

First, we ask: how well can the model predict for advertisers it has already seen? This addresses the model's utility in a stable environment with a known set of advertisers. To answer this, we employ a stratified $k$-fold cross-validation that trains the model on a subset of observations from every advertiser and tests it on a held-out subset from those same advertisers, thereby measuring

Figure 4: PIE model performance by Prospecting vs. Retargeting




*Notes:* This figure compares the out-of-sample $R^2$ to predict ICPD ($\hat{\psi}_r$), decomposed across campaigns that either primarily target new customers ("Prospecting") or existing customers ("Retargeting"). The bars represent progressive additions of features to the PIE model, from PIE(Pre) (using only pre-determined features) to PIE(Full) (using all features). Raw LCC-7D serves as a benchmark that divides the raw last-click 7-day conversion count by campaign spend and treats the result as if it were ICPD, with no model applied. The $R^2$ value is shown above each bar, and error bars represent one standard deviation from the mean $R^2$, estimated via bootstrap.

its interpolative accuracy.[23]

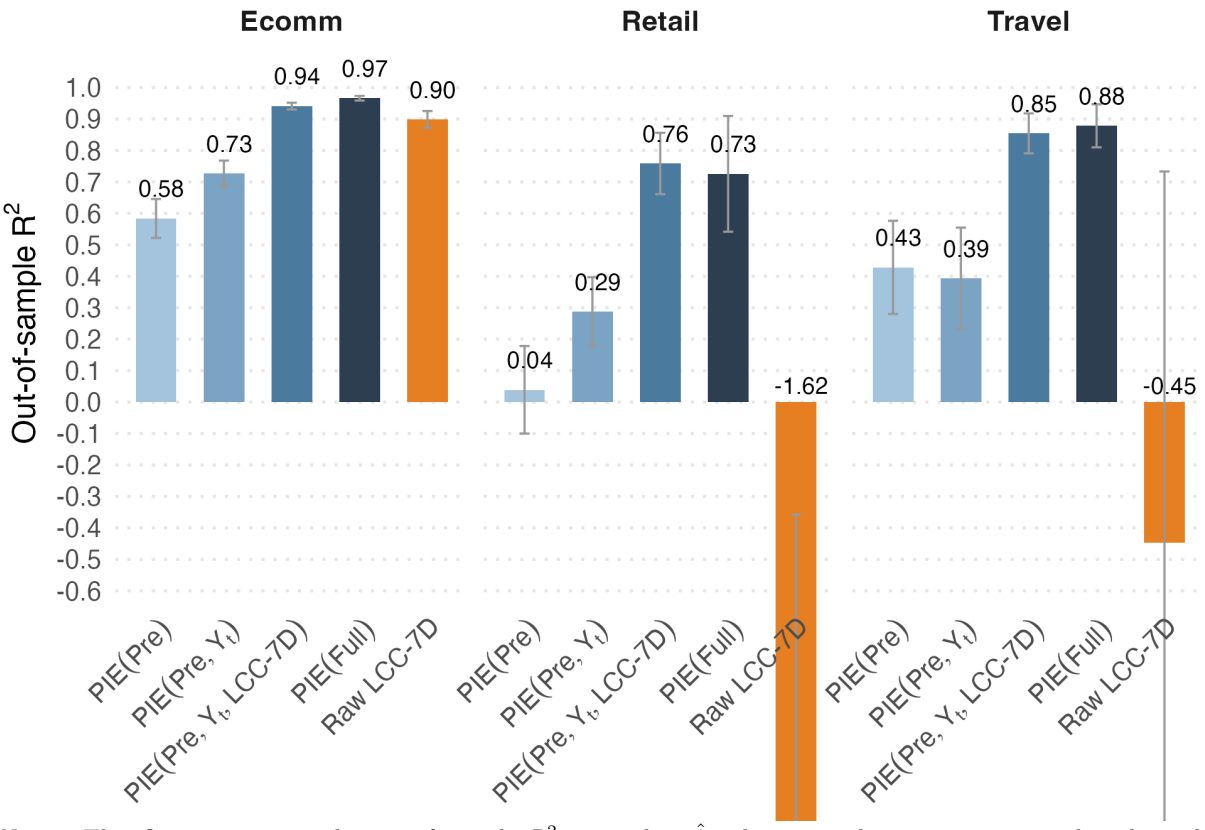
Second, we ask: how well can the model generalize to new advertisers it has never seen before? This addresses the "cold-start" problem, a common challenge in marketing applications. To answer this more stringent question, we use a group $k$-fold cross-validation in which the holdout sets are composed of entirely unseen advertisers, meaning the model is trained on one cohort of advertisers and evaluated on its ability to predict for a completely separate cohort.

To implement these exercises within our data, we focus on the subset of advertisers with at least two RCTs, which yields a slightly smaller sample of 2,103 observations. In each exercise, we estimate the sampling variance using a non-parametric bootstrap across advertisers, sampling with replacement and performing each cross-validation procedure in its entirety for each bootstrap iteration.

These results are presented in Figure 7. On the left, we repeat the performance obtained by using PIE(Full) based on the full sample of 2,226 RCTs from Figure 2. The second condition, "2+ RCTs", presents the predictive evaluation from 10-fold cross-validation using the narrower sample

[23] At the extreme, we could use a leave-one-RCT-out analysis to estimate the same quantity, but this is a relatively costly operation and in some testing we did not observe significant gains from doing so.

Figure 5: PIE model performance by industry vertical



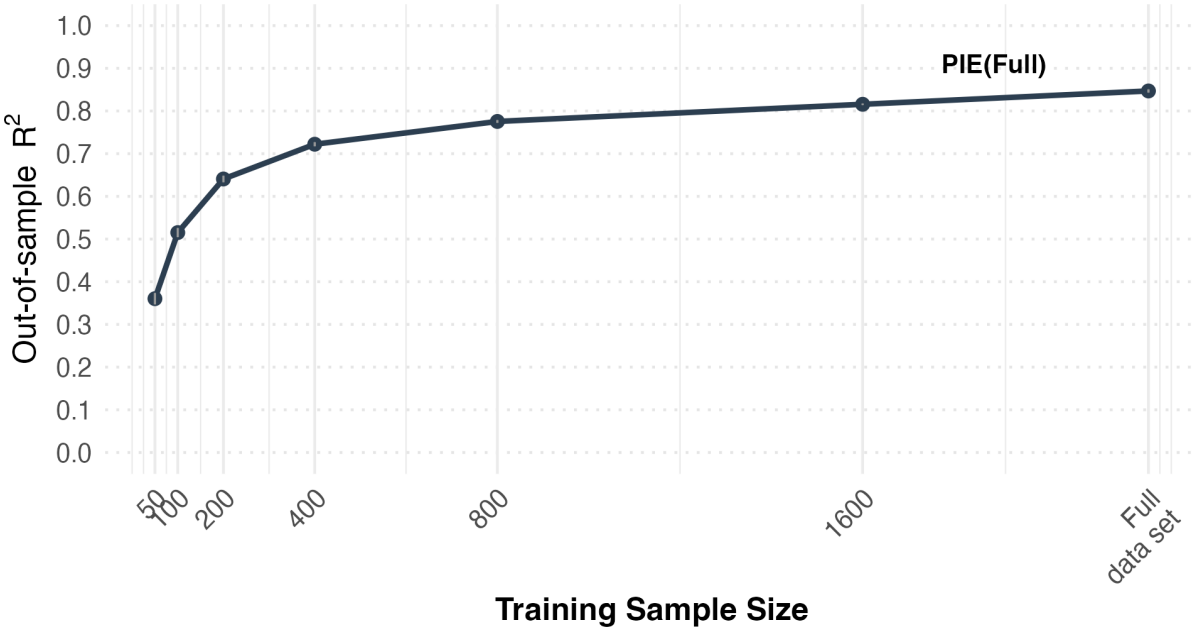

*Notes:* This figure compares the out-of-sample $R^2$ to predict $\hat{\psi}_r$, decomposed across campaigns based on the product's industry vertical: e-commerce ("Ecomm"), retail, or travel. The bars represent progressive additions of features to the PIE model, from PIE(Pre) (using only pre-determined features) to PIE(Full) (using all features). Raw LCC-7D serves as a benchmark that divides the raw last-click 7-day conversion count by campaign spend and treats the result as if it were ICPD, with no model applied. The $R^2$ value is shown above each bar, and error bars represent one standard deviation from the mean $R^2$, estimated via bootstrap. The $R^2$ values for Raw LCC-7D in Retail and Travel are substantially negative, such that we have truncated the plot to maintain a readable scale for the positive $R^2$ values. The $R^2$ values shown by the text label (e.g., "$-1.62$") represent the true, untruncated values.

of advertisers with two or more RCTs. Thus, the performance of the PIE model on this slightly smaller sample of 2,103 RCTs is comparable to its performance on the full sample of 2,226 RCTs. Consistent with this finding, the raw last-click 7-day metric yields an $R^2$ that is 0.70 lower than the full PIE model.

PIE achieves the same performance as in the baseline when interpolating for "Existing Advertisers." When extrapolating to "New Advertisers", the $R^2$ of the PIE model drops from 0.87 to 0.72 relative to the "Existing Advertisers" case. A reduction in predictive accuracy is expected when a model is used for extrapolation. It is difficult to assess whether this increase in error is "large," in an absolute sense, since this depends on how different the new advertisers are relative to the training sample.

One useful point of comparison is the last-click 7-day metric. This metric yields the same

Figure 6: PIE model performance as a function of training sample size.




*Notes:* This figure illustrates the relationship between training sample size (number of RCTs) and the out-of-sample $R^2$ for the PIE(Full) model. The $R^2$ values were obtained by repeatedly training the PIE(Full) model on random subsamples of increasing size and evaluating its predictive performance.

$R^2 = 0.17$ regardless of whether the prediction is for new or existing advertisers, and performs substantially worse than PIE. This implies that an advertiser who had never run an RCT on the platform would benefit from evaluating their non-RCT campaigns using the full PIE model instead of the campaign's last-click 7-day conversions.

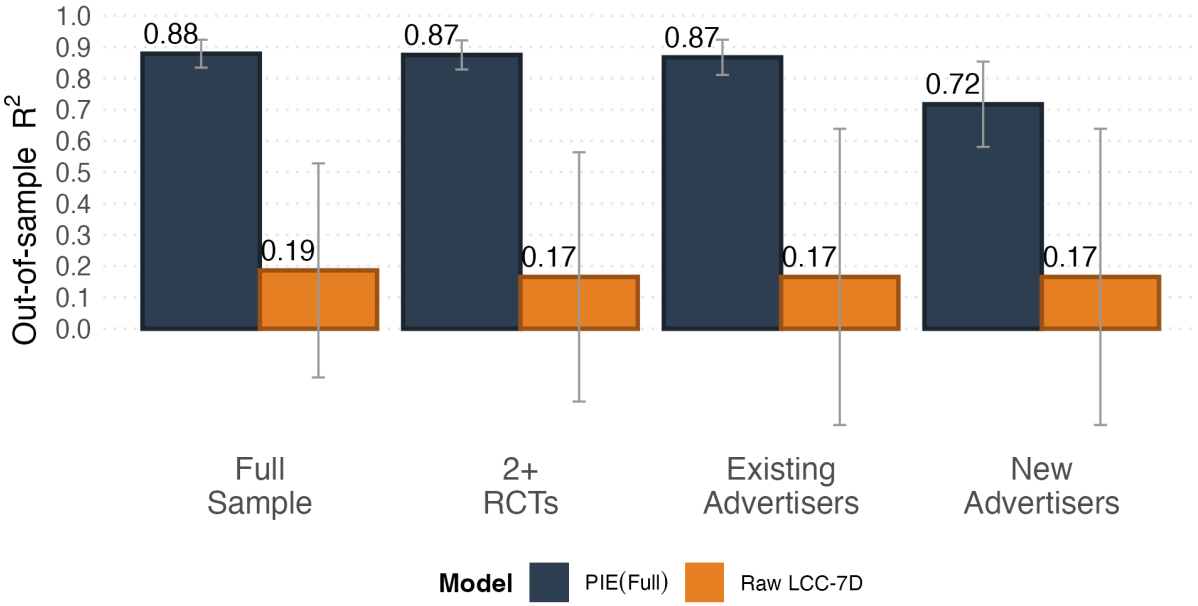
### 5.3 Predictive Accuracy for New Campaign Segments

The advertiser-level cross-validation in the preceding subsection tells us whether the model can generalize across *firms*. In practice, however, a platform must also predict outcomes across *segments* of ad campaigns that differ in creative strategy, campaign objectives, geography, or product vertical. To probe this second dimension of robustness, we conduct a hold-out-one-level experiment that asks a stricter question: How much accuracy is lost when the model must extrapolate from other segments to a focal segment it has never seen in training?

Although the advertiser-level cross-validation removes the target firm from the training set, the model may still learn from closely related advertisers. For example, if the training folds include Nike and Reebok, the model can transfer information related to their campaign features to predict effects for Adidas, even if Adidas is excluded from the training set. This overlap softens the cold-start problem.

The segment-level experiment that follows is intentionally stricter. When we hold out one categorical level (e.g., advertisers belonging to the retail vertical), *all* advertisers belonging to that level are excluded from training. The model must therefore extrapolate across both firms and

Figure 7: PIE Performance for Existing and New Advertisers



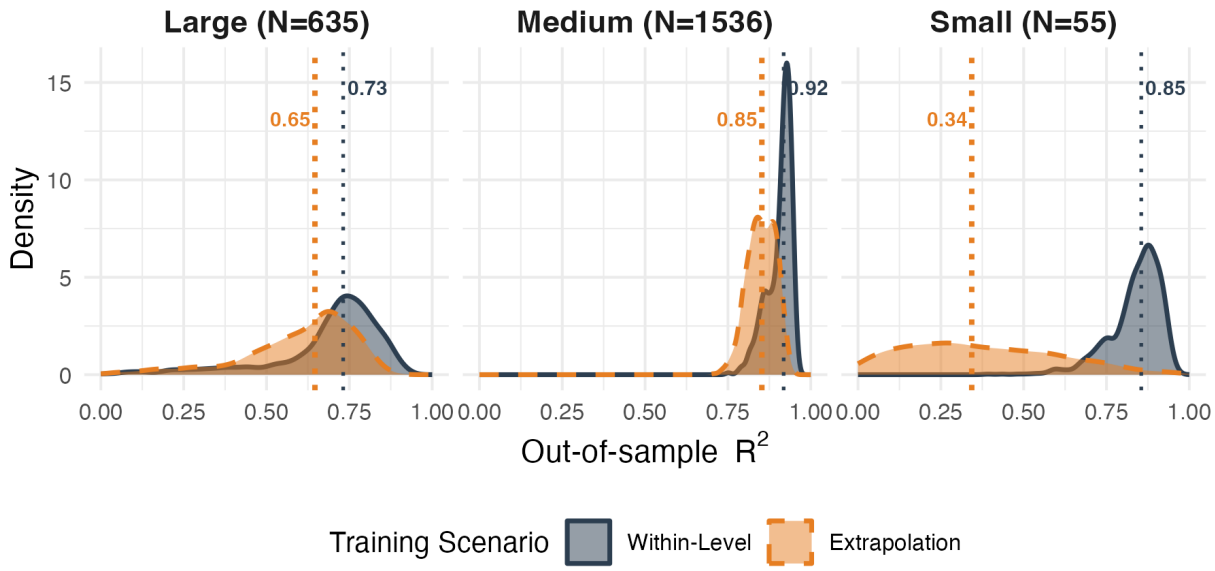

*Notes:* This figure presents the out-of-sample $R^2$ for the PIE(Full) model and the Raw LCC-7D benchmark across four conditions. The first two groups are based on simple $k$-fold cross-validation, using the Full Sample (all 2,226 RCTs) and a restricted sample (2+ RCTs) used for the subsequent analyses, respectively. The remaining two groups of bars test the model's robustness using specialized cross-validation approaches: Existing Advertisers measures interpolative accuracy using stratified $k$-fold CV (training and testing on different observations from the same advertisers). New Advertisers measures extrapolative accuracy using group $k$-fold CV (training on one cohort of advertisers and testing on an entirely separate, unseen cohort). Error bars represent one standard deviation from the mean $R^2$.

segments, providing a sharper test of its ability to transfer knowledge to genuinely novel campaign contexts. By pairing two otherwise identical models—one trained inside the focal level and the other trained entirely outside it—we obtain an empirical distribution of "extrapolation penalties" that complements the advertiser-level metrics reported earlier.

We conduct this exercise using the following categorical features: advertiser size (small, medium, large), campaign year (using 2019 campaigns to predict 2020 campaigns), custom audience (broad or narrow), conversion optimization (offsite or onsite), prospecting or retargeting campaigns, and advertiser vertical (e-commerce, retail, and travel).

Operationally, for each level $\ell$ of a chosen campaign feature (e.g., prospecting campaigns), we first randomly divide all observations belonging to that level into two equal halves, each of size $R_\ell/2$. The first half is used to estimate a *within-level* PIE model, while the second half is retained as a hold-out set unseen during training. A second model, the *extrapolation* PIE model, is then trained on a sample of identical size $R_\ell/2$ drawn with replacement from the pool of observations outside level $\ell$ (e.g., retargeting campaigns). Both models are evaluated on the *same* hold-out set, such that any performance gap reflects only the cost of relying on observations from beyond the

Figure 8: Extrapolation performance by advertiser size




*Notes:* Empirical distributions of out-of-sample $R^2$ from the segment-level hold-out experiment, shown here for advertiser size. Each panel corresponds to one advertiser-size group (*small*, *medium*, *large*), with separate density curves for models trained within the focal group (red) and for models trained by extrapolating from the remaining groups (blue). The distributions are based on 1,000 repeated random partitions of the data, providing a view of how predictive accuracy varies across resamples under the within- and extrapolation-training conditions. Vertical dotted lines indicate the median $R^2$ within each training scenario.

segment's focal level.[24] This entire split-train-test cycle is repeated 1,000 times with fresh random partitions, producing a distribution of $R^2$ for both scenarios at every level $\ell$.

Figure 8 visualizes these densities according to the focal segments defined by advertiser size, with figures for the other segmentation variables located in Appendix 4. We rely on a classification scheme that categorizes advertisers as "small," "medium," and "large." Densities with lower within-scenario variance indicate that the model achieves fairly consistent within-segment predictive performance.

The extent to which the density for the extrapolation scenario (blue) is shifted to the left indicates the additional error induced by extrapolating from a model trained on data that completely lack the target level. In the left subfigure of Figure 8, we observe more consistency and little relative extrapolative error when the PIE model uses non-large advertisers to predict large advertisers. The penalty remains similarly low when we shift to predicting medium advertisers, which constitute the majority of observations in our sample. However, the extrapolative error increases sharply when we use a model trained on large and medium advertisers to predict ad effects for small advertisers, of which we have only 55 observations.

[24] For example, consider the segmentation variable "advertiser size" $\in \{\text{small}, \text{medium}, \text{large}\}$. If the focal level is $\ell = \text{large}$, then we first split all campaigns with large advertisers into two groups, training on one and testing on the other, to produce our within-model evaluation. Next, we train the extrapolation model using an equal-sized training sample drawn randomly from the small and medium advertisers, and evaluate it using the same held-out test sample as with the within-model.

Table 1 on page 29 reports the difference in the median out-of-sample $R^2$ across training scenarios for all the segmentation variables and their levels (see Appendix 4 for additional plots). Across the segmentation variables, PIE predictions for extrapolations yield between 0.01 and 0.25 lower $R^2$ than their within-level PIE predictions.

Table 1: Out-of-sample $R^2$ performance: within-level versus extrapolation scenarios.

| Segmentation Variable | Within-Level $R^2$ | Extrapolation $R^2$ | Difference |
|---|---|---|---|
| **Conversion optimization** | **0.837** | **0.784** | **-0.053** |
| Opt Onsite Conversion | 0.854 | 0.824 | -0.031 |
| Opt Offsite Conversion | 0.813 | 0.693 | -0.120 |
| **Advertiser size** | **0.852** | **0.674** | **-0.178** |
| Large | 0.731 | 0.639 | -0.092 |
| Medium | 0.918 | 0.852 | -0.066 |
| Small | 0.855 | 0.303 | -0.552 |
| **RCT Year** | **0.812** | **0.599** | **-0.213** |
| 2020 | 0.812 | 0.599 | -0.213 |
| **Prospecting vs. retargeting** | **0.824** | **0.767** | **-0.057** |
| Retargeting | 0.837 | 0.696 | -0.141 |
| Prospecting | 0.806 | 0.796 | -0.009 |
| **Vertical** | **0.825** | **0.691** | **-0.134** |
| Ecomm | 0.939 | 0.883 | -0.056 |
| Retail | 0.724 | 0.622 | -0.102 |
| Travel | 0.807 | 0.556 | -0.250 |
| **Custom audience** | **0.706** | **0.559** | **-0.147** |
| Broad Audience | 0.736 | 0.612 | -0.124 |
| Narrow Audience | 0.587 | 0.434 | -0.152 |

*Notes:* Each row reports the median out-of-sample $R^2$ for the within-level and extrapolation training scenarios, along with their difference. Bolded values in the header rows represent weighted medians across levels within each segmentation variable.

# 6 Do Predictive Models Lead to the Same Decisions?

So far, we have focused on whether the ad effects estimates obtained from predictive models come close to those from RCTs. In practice, however, the estimates serve as a means to an end, which is to inform a managerial decision. As pointed out by Fernández-Loría and Provost (2022), the task of estimating a causal effect is distinct from the task of making an optimal decision. An estimated effect with lower prediction error does not guarantee it will lead to better decisions, because the costs of different errors may be asymmetric and the magnitude of the error may matter differently depending on whether the estimate falls near a decision threshold. Therefore, to assess the practical value of PIE, we must evaluate it within a decision-making framework. What if a manager would make the same campaign decision regardless of whether they used a predictive model or an RCT? If this is the case, the practical implications of predictive models would be lessened despite their potential inaccuracy.

Our goal in this section is to assess the extent to which an advertiser using either last-click attribution or the PIE model would reach different conclusions about campaign effectiveness com-

pared with using an RCT. We first formalize the advertiser's decision problem and describe how the use of different measurement methodologies can lead to different conclusions about a campaign's success. We then quantify these differences by estimating the probability that the RCT-based and model-based decisions would *disagree*.

### 6.1 Disagreement Probabilities

How should we think about managerial decisions related to advertising campaigns? We adopt a perspective common in industry practice: marketers evaluate campaign performance based on incremental conversions per dollar (ICPD), as defined in Section 4.2.

We assume that a manager considers a campaign successful if its ICPD exceeds a threshold, $\text{ICPD}^*$. Otherwise, the campaign is deemed unsuccessful, and the manager would prefer to reallocate those marketing funds elsewhere. The threshold can be interpreted as the inverse of the break-even profit-per-conversion that would justify the campaign's cost. The key question is whether the manager would reach the same judgment (success or failure) when using model-based estimates as when using RCT-based ones.[25]

To illustrate, consider a manager who sets an advertising budget of \$100,000 and an ICPD threshold of $\text{ICPD}^* = 0.01$, corresponding to one incremental conversion per \$100 spent. If an RCT shows that the campaign generated 500 incremental conversions, the observed $\text{ICPD}_r = 500/100{,}000 = 0.005$, which falls short of the threshold. The manager would thus conclude the campaign failed.

Now suppose instead that the manager relied on a model that estimated 2,000 incremental conversions, implying $\text{ICPD}_r^{\text{pred}} = 0.02$. Based on this model, the campaign would appear successful. The RCT and model therefore lead to opposing judgments (a disagreement).

This particular case represents a false positive, or Type I error: the model concludes that the campaign was successful when the RCT indicates otherwise. If the opposite occurred—the RCT judged the campaign successful but the model did not—the disagreement would represent a false negative, or Type II error. When both methods classify the campaign on the same side of the threshold, we say they agree.

We estimate the disagreement probability between the RCT and two models, the PIE(Full) and last-click 7-day. A high disagreement probability implies that model-based evaluations would often lead advertisers to make different campaign decisions than they would under the RCT benchmark, potentially misallocating spend despite small average estimation errors. Conversely, a low disagreement probability means that, even if model estimates are biased, advertisers would typically reach the same go/no-go decisions as they would with an RCT, making the model a practically reliable substitute for experimental measurement.

[25] The ICPD is simply the inverse of a "Cost per X" metric. Many marketers evaluate campaign success using measures such as Cost per Conversion (CPC), Cost per Acquisition (CPA), Cost per Completed View (CPCV), or Cost per Engagement (CPE) (Xaxis, 2018). These metrics relate campaign costs to outcomes and are typically interpreted as if they reflect incremental performance, even though they are often based on non-causal measures. See `https://growhackscale.com/glossary/cost-per-conversion-metric`, accessed October 12, 2025.

For a given value of ICPD$^*$, we calculate the disagreement probability over the set of campaigns as the sum of the Type I and Type II error probabilities:

$$\text{Disagreement}(\text{ICPD}^*) = \frac{1}{R}\sum_{r\in\mathcal{R}} \underbrace{\mathbb{1}\{\text{ICPD}_r \leq \text{ICPD}^*\} \cdot \mathbb{1}\{\text{ICPD}_r^{\text{pred}} > \text{ICPD}^*\}}_{\text{Type I (false positive)}} + \underbrace{\mathbb{1}\{\text{ICPD}_r > \text{ICPD}^*\} \cdot \mathbb{1}\{\text{ICPD}_r^{\text{pred}} \leq \text{ICPD}^*\}}_{\text{Type II (false negative)}}.$$

Because our interest lies in whether a manager would make the same campaign-level decision when relying on model-based estimates rather than the RCT benchmark, each campaign represents a single decision instance. We therefore compute the disagreement probability by giving every campaign equal weight.

To define the relevant range of ICPD$^*$ thresholds, we normalize the ICPD values for campaigns according to the median ICPD among campaigns in the same industry vertical and with outcomes at the same event funnel position. We do so because campaigns optimizing for upper-funnel events (e.g., page views) typically exhibit higher ICPD than those optimizing for lower-funnel conversions (e.g., purchases), and there is natural variation in ICPD across verticals. We examine thresholds in the range of 50% to 150% of each median, helping to ensure the thresholds we consider are economically relevant within each campaign segment.

We implement this approach using a non-parametric bootstrap, sampling observations with replacement and then training the PIE(Full) model on each sample (using 10-fold cross-validation). Then we estimate the disagreement probabilities point-wise at each ICPD$^*$, with the median normalization computed separately for each bootstrap replication. This process is repeated 1,000 times.

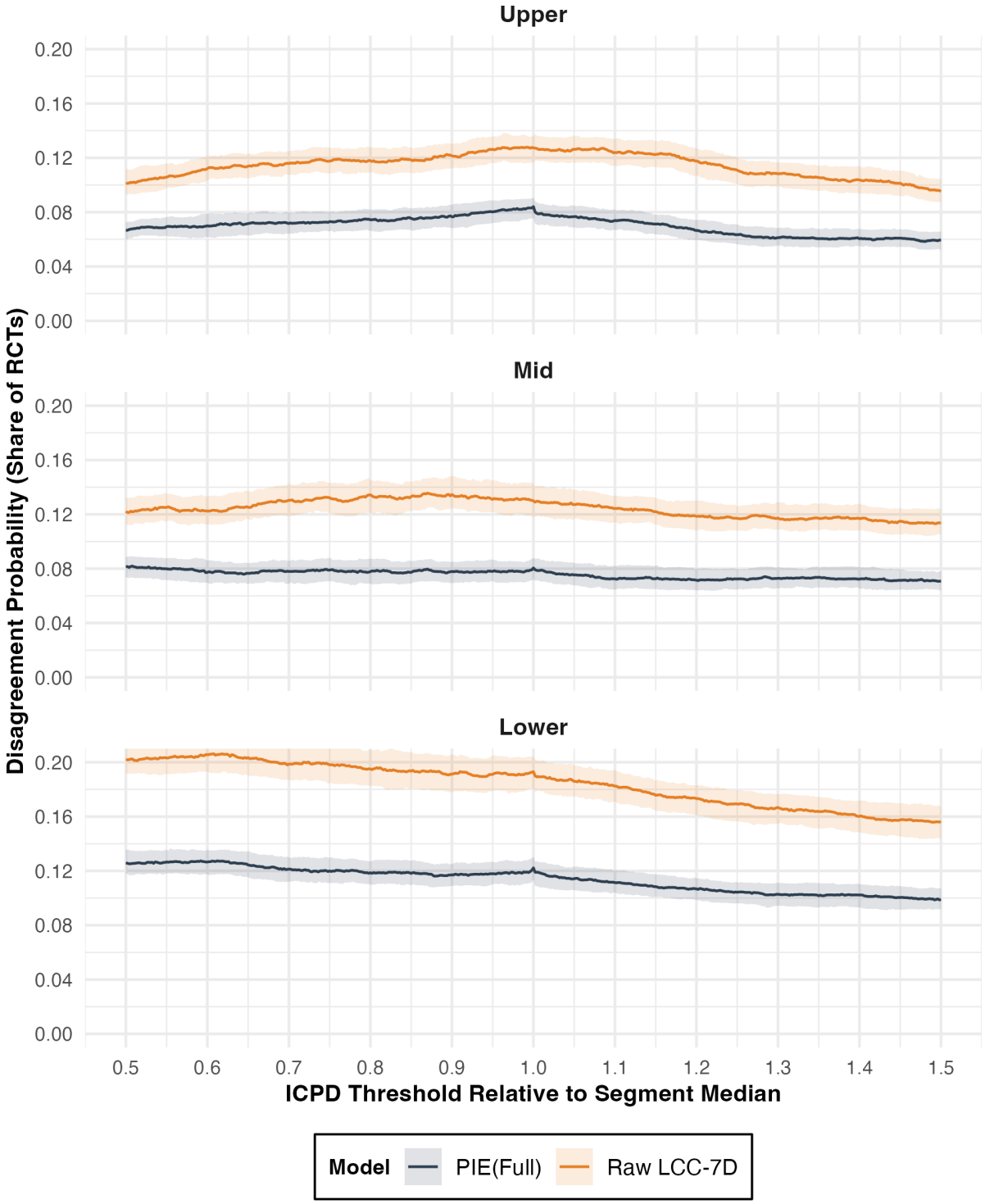
### 6.2 Results

Figure 9 plots the median disagreement probabilities by funnel level when using either the PIE(Full) model vs. RCT (dark blue) or the raw last-click 7-day vs. RCT (orange), with the shaded regions corresponding to the inter-quartile range. The value of one on the x-axis corresponds to the median value within any given campaign's segment (vertical-funnel level).

To interpret the figure, suppose that an advertiser sets a threshold of ICPD$^* = 1$. At such a threshold, the ICPD derived from a last-click 7-day attribution metric would "disagree" with the ICPD from an RCT in approximately 19.0% of the lower-funnel campaigns, compared with approximately 12% when using the PIE model. This finding for last-click attribution should not be generalized beyond the present setting, as the degree of attribution bias is likely platform-specific. In search advertising, clicks arise from latent purchase-intent prompting a query. In contrast, display ad clicks may reflect interest-based targeting and may therefore exhibit a different relationship between click behavior and true incremental effects.

Figure 10 decomposes the disagreement curves in Figure 9 into their Type I and II error components (omitting the bootstrapped confidence bands for clarity). The key observation is the

Figure 9: Disagreement Probability




*Notes:* Plots the share of RCTs for which an advertiser's decision about campaign success would differ when using model-based estimates of incremental conversions per dollar (ICPD) instead of the RCT benchmark, separately by event funnel. Decisions are based on whether ICPD exceeds a threshold defined relative to each segment's (event-funnel and vertical) median ICPD. The curves report the median disagreement probability across bootstrap repetitions for the PIE(Full) model (dark blue) and the raw last-click conversion 7-day metric (orange). Shaded regions denote the interquartile range across bootstrap repetitions.

asymmetry in error rates when relying on last-click 7-day conversions. The dashed orange line for last-click 7-day shows a lower Type II error rate but a substantially higher Type I error rate. This pattern arises because last-click 7-day metrics tend to be upwardly biased: by capturing conversions that occurred up to seven days after the impression and click-through, they systematically inflate ICPD and compress variation across campaigns. As a result, last-click rarely classifies a genuinely successful campaign as unsuccessful (low Type II error), but it frequently classifies unsuccessful campaigns as successful (high Type I error).

In contrast, the PIE model achieves low and balanced probabilities on both types of errors. If an advertiser is primarily focused on minimizing Type II error, then either the last-click 7-day metrics or PIE model perform similarly. However, if an advertiser cares similarly about both error types, then the PIE model outperforms last-click.

Note that a manager may have different costs for committing either type of error. A false positive—incorrectly concluding that the campaign was a success—might lead the manager to implement future ad campaigns that are similarly unsuccessful, potentially wasting a significant amount of funds over time. A false negative might lead to the opposite outcome, causing an advertiser to hold back on further advertising spending even though it would have been the right decision. In this paper, we cannot assess the relative magnitudes of these costs. However, this information could be incorporated into the decision problem if the advertiser has access to such information.

# 7 Conclusion

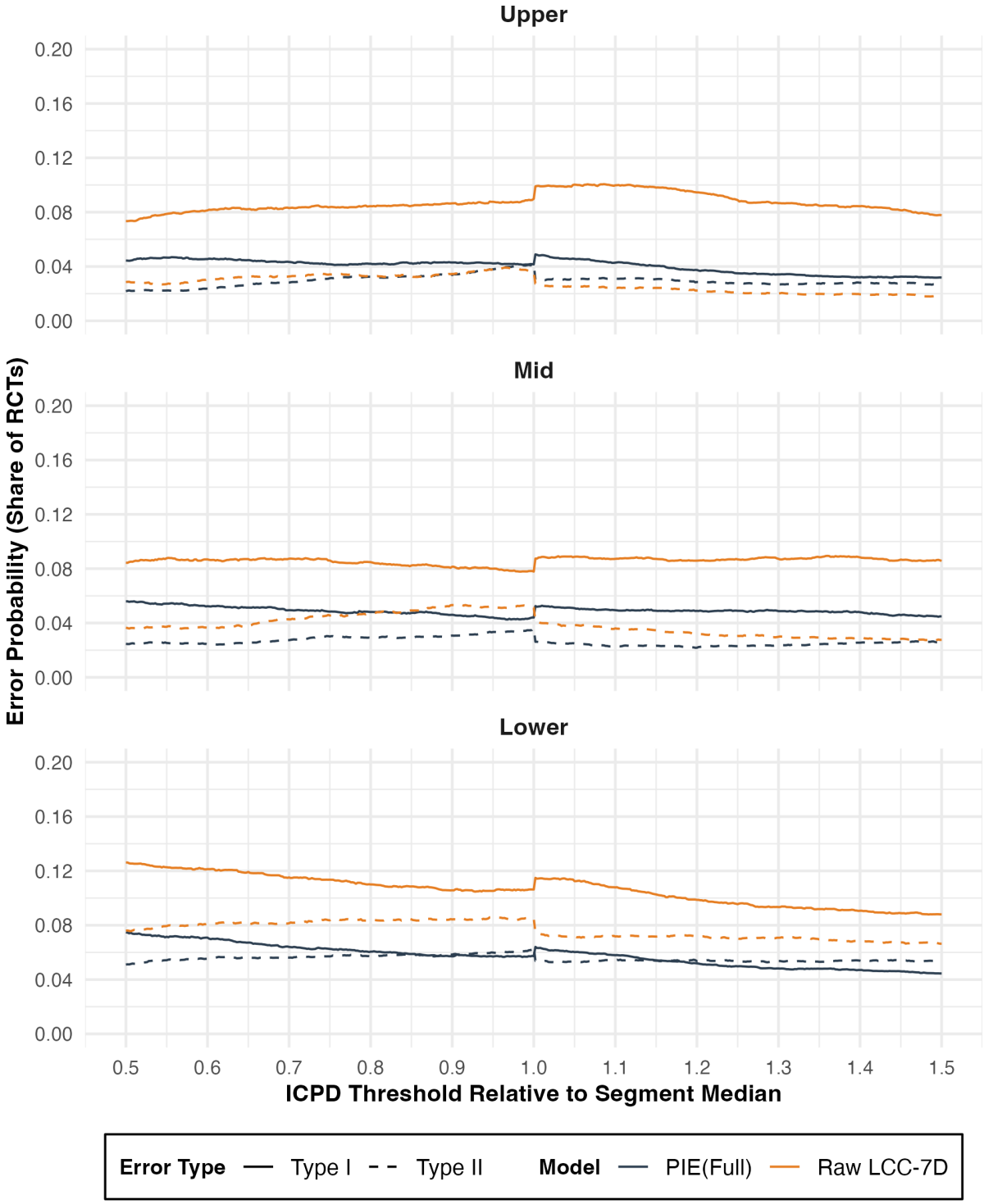
Advertisers and platforms need credible measures of incrementality, yet always-on RCTs are computationally costly to run at scale and observational proxies can be biased.

We propose Predicted Incrementality by Experimentation (PIE), a method to scale RCT causal effects to non-RCT campaigns. PIE uses a set of RCTs to learn a mapping from pre-determined (e.g., campaign configuration settings) and post-determined features (e.g., average outcomes, exposure and click behaviors, attribution counts) to the causal quantity of interest, and then applies that mapping to campaigns not run as RCTs. We view PIE first and foremost as a prediction, or supervised learning, problem. The emphasis is on accuracy in out-of-sample prediction, not identification of underlying mechanisms or user-level causal parameters.

Our framework clarifies why post-determined features—realized aggregates from the test group that most platforms already log—contain useful signal about incrementality. The same forces that shape causal effects, such as consumers' organic conversion rate and selection into exposure, also shape these post-determined features. PIE leverages this linkage by learning how multiple aggregate metrics can be combined to predict the causal estimand.

Our primary contribution is to reframe non-RCT measurement as campaign-level prediction anchored by experiments and to formalize why post-determined features contain signal about incrementality. Because PIE relies only on test-group aggregates, it is feasible to deploy at scale with standard safeguards.

Figure 10: Disagreement Probability by Type I and II Errors




*Notes:* Decomposes the disagreement probability from Figure 9 into Type I errors (solid lines: false positives, where the model predicts ICPD above threshold but the RCT benchmark is below) and Type II errors (dashed lines: false negatives, where the model predicts ICPD below threshold but the RCT benchmark is above), separately by event funnel position. Thresholds are defined relative to each segment's median ICPD. The curves report the median error probability across bootstrap repetitions for the PIE(Full) model (dark blue) and the last-click conversion 7-day metric (orange).

We present a proof-of-concept application of PIE using a sample of 2,226 ad experiments at Meta. PIE explains approximately 90% of the out-of-sample variation in RCT effects and improves over common attribution benchmarks. PIE generalizes well within advertisers and, as expected, less well across segments that differ markedly from the training pool.

PIE is a general framework: causal effects can be mapped from RCTs to non-RCTs beyond the advertising domain, provided similar post-determined features that encode relevant structural parameters are available.

Because PIE's accuracy depends on representativeness and overlap, an ad platform implementing PIE should curate a donor pool of RCTs that is both large and intentionally diverse across campaign characteristics (e.g., verticals, objectives, audience strategies, spend tiers, funnel events). In practice, this implies seeding RCTs where coverage is sparse (e.g., new objectives or formats), and excluding idiosyncratic tests that reflect one-off engineering configurations. PIE relies on the invariance of the relationship between features ($\mathbf{X}_r$) and causal targets ($\hat{\psi}_r$), so a platform must assess whether the distribution of features appears to shift over time and also ensure temporal coverage (rolling inclusion of recent campaigns).

Finally, PIE is agnostic to precisely how the RCT outcomes are generated. We view platform-initiated experiments as particularly promising for generating PIE training data. In a "shadow mode" implementation, the platform continuously runs backend RCTs, invisible to advertisers, to fill gaps in the training pool. This helps ensure a steady supply of representative RCTs for model training. When the causal effects (labels) are noisier, they can be pooled with inverse-variance (meta-analytic) weights or stabilized via Bayesian hierarchical/shrinkage models and then fed into a supervised learning model. Choosing among these designs involves familiar trade-offs: the opportunity cost of withholding impressions, the engineering burden of managing a smaller number of shadow tests versus many smaller experiments, and the precision of each causal effect versus the representativeness of the donor pool. We view these as product and operations decisions rather than conceptual changes to PIE, and we leave a full treatment to future work.

## References


Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics 130*(3), 1117–1165.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might Be Ineffective. *Journal of Marketing Research 55*, 80–98.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2025). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *The Review of Economic Studies*. Forthcoming.

Athey, S. and G. W. Imbens (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Fernández-Loría, C. and F. Provost (2022). Causal Decision Making and Causal Effect Estimation Are Not the Same...and Why It Matters. *INFORMS Journal on Data Science 1*(1), 4–16.

Google (2024). *Modern Measurement Playbook: How to use media effectiveness measurement to make better business decisions*. Google. 2024 edition; Primary Author: Ana Carreira Vidal, EMEA Media Effectiveness Lead.

Gordon, B. R., K. Jerath, Z. Katona, S. Narayanan, J. Shin, and K. C. Wilbur (2021). Inefficiencies in digital advertising markets. *Journal of Marketing 85*(1), 7–25.

Gordon, B. R., R. Moakler, and F. Zettelmeyer (2023). Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement. *Marketing Science 42*(4), 768–793.

Gordon, B. R., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2019). A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. *Marketing Science 38*(2), 193–225.

Gui, G., H. Nair, and F. Niu (2022). Auction Throttling and Causal Inference of Online Advertising Effects. Technical report, arXiv. arXiv:2112.15155.

Heckman, J. J. and E. Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica 73*(3), 669–738.

Hernán, M. A. and J. M. Robins (2020). *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC. Freely available at `https://miguelhernan.org/whatifbook`.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005). Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations. *Journal of Econometrics 125*(1–2), 241–270.

Huang, T.-W., E. Ascarza, and A. Israeli (2024). Incrementality Prediction: Synergizing Past Experiments for Intervention Personalization. Working paper, SSRN. Available at SSRN: 4859809.

Imbens, G. W. and J. D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (1st ed.). Cambridge University Press.

Jin, Y., Y. Wang, Y. Sun, D. Chan, and J. Koehler (2017). Bayesian methods for media mix modeling with carryover and shape effects. Technical report, Google Inc.

Johnson, G. A. (2023). Inferno: A Guide to Field Experiments in Online Display Advertising. *Journal of Economics and Management Strategy 32*(3), 469–490.

Johnson, G. A., R. A. Lewis, and E. I. Nubbemeyer (2017). Ghost Ads: Improving the Economics of Measuring Ad Effectiveness. *Journal of Marketing Research 54*(6), 867–884.

Lewis, R., J. Rao, and D. Reiley (2011). Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising. In *Proceedings of the 20th International Conference on World Wide Web*, pp. 157–66. Association for Computing Machines.

Lewis, R. and D. Reiley (2014). Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on Yahoo! *Quantitative Marketing and Economics 12*(3), 235–266.

Lewis, R., F. Zettelmeyer, B. R. Gordon, C. Garib, J. Hermle, M. Perry, H. Romero, and G. Schnaidt (2025). Amazon Ads Multi-Touch Attribution. https://arxiv.org/abs/2508.08209.

Lewis, R. A. and J. M. Rao (2015). The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics 130*(4), 1941–1973.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

Pearl, J. and E. Bareinboim (2011, Aug.). Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence 25*(1), 247–254.

Rafieian, O. and H. Yoganarasimhan (2022). Variety effects in mobile advertising. *Journal of Marketing Research 59*(4), 718–738.

Runge, J., H. Patter, and I. Skokan (2023, March). A New Gold Standard for Digital Ad Measurement? Harvard Business Review.

Runge, J., I. Skokan, G. Zhou, and K. Pauwels (2025). Packaging up media mix modeling: An introduction to robyn's open-source approach.

Rzepakowski, P. and S. Jaroszewicz (2012). Uplift modeling in direct marketing. *Journal of Telecommunications & Information Technology 2012*, 43.

Schiff, A. (2020). Facebook Conversion Lift Measurement Issue Goes Undetected For 12 Months. `https://www.adexchanger.com/platforms/facebook-conversion-lift-measurement-issue-goes-undetected-for-12-months/`. adexchanger.com.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science 25*(3), 289–310.

Stanley, T. D. and H. Doucouliagos (2012). *Meta-Regression Analysis in Economics and Business*. New York, NY: Routledge.

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association 18*(6), 3045–3089.

Wager, S. and S. Athey (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Waisman, C. and B. R. Gordon (2025). Multicell Experiments for Marginal Treatment Effect Estimation of Digital Ads. *Management Science 71*(9), 7970–7989.

Xaxis (2018). Measuring Success: A Global Study of How Marketers Prove the Value of their Digital Media Investments. Technical report, Xaxis.

Yang, J., D. Eckles, P. Dhillon, and S. Aral (2024). Targeting for Long-Term Outcomes. *Management Science 70*(6), 3841–3855.

# Appendix 1 Notation

This table summarizes the main notation used throughout the paper.

Table A-1: Key Notation

| Symbol | Meaning |
|---|---|
| $r$ | Campaign/RCT index |
| $i$ | User index within campaign |
| $R$ | Total number of campaigns/RCTs |
| $N_r$ | Number of users in RCT $r$ |
| $N_{tr}$ | Number of users in RCT $r$ in the test group |
| $N_{cr}$ | Number of users in RCT $r$ in the control group |
| $Z_i^r$ | Random assignment indicator (1 = test, 0 = control) |
| $D_i^r$ | Exposure indicator (1 = exposed, 0 = not exposed) |
| $C_i^r$ | Click indicator (1 = clicked, 0 = did not click) |
| $U_i^r$ | User- and RCT-specific unobservable |
| $\alpha_r$ | Average baseline outcome (without exposure) in RCT $r$ |
| $\tau_r$ | Average treatment effect of exposure in RCT $r$ |
| $\gamma_r$ | Effect of user unobservable on baseline outcomes |
| $\delta_r$ | Causal effect of clicking on outcomes |
| $\phi_r$ | Exposure rate in test group of RCT $r$ |
| $\overline{Y}_{tr}$ | Average outcome in test group of RCT $r$ |
| $\overline{Y}_{cr}$ | Average outcome in control group of RCT $r$ |
| $\psi_r$ | Target causal estimand for PIE prediction |
| $\mathbf{X}_r^{\text{pre}}$ | Pre-determined features for RCT $r$ |
| $\mathbf{X}_r^{\text{post}}$ | Post-determined features for RCT $r$ |
| $\text{ICPD}_r$ | Incremental conversions per dollar for RCT $r$ |
| $\text{LCC}_r(w)$ | Last-click conversions with attribution window $w$ |

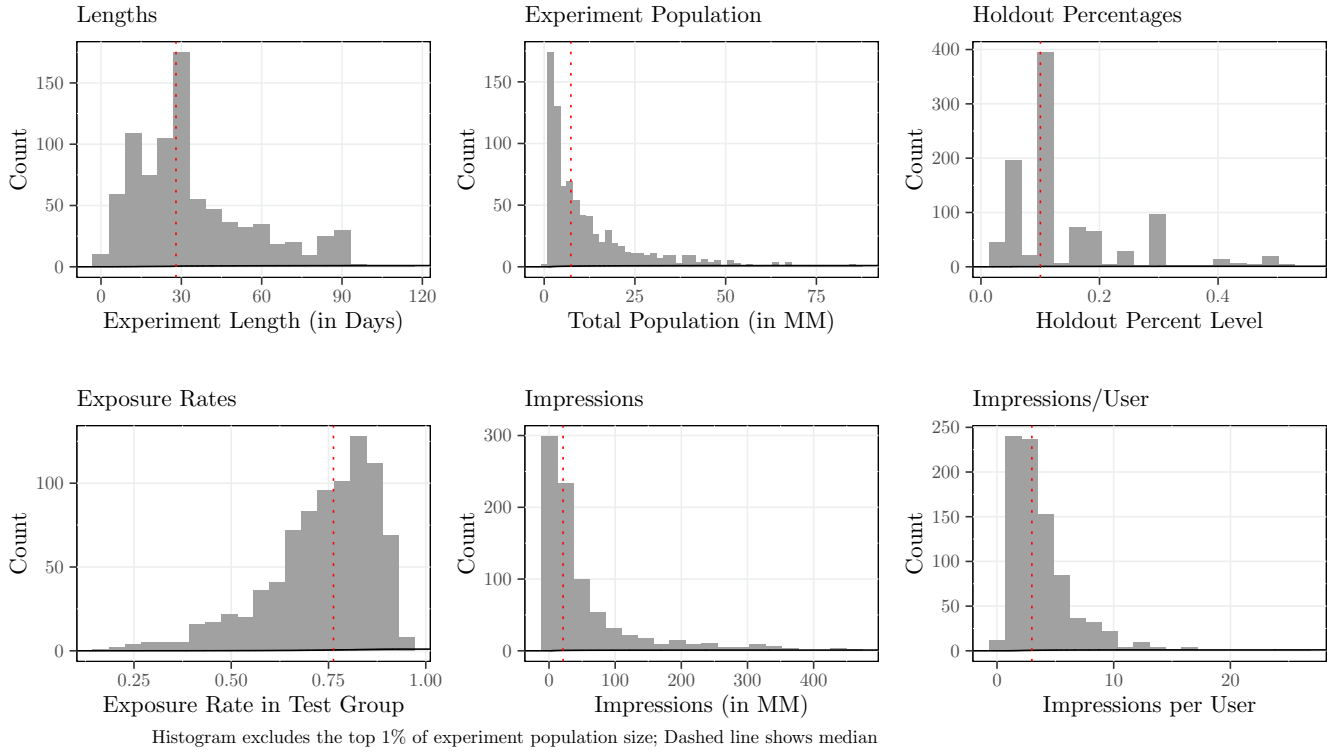
# Appendix 2 Data Details

This section provides additional details on the dataset of Meta ad experiments. Recall that we selected experiments with at least one million users assigned to test and retained outcome metrics with at least 5,000 conversions in the test group. These choices balance representativeness with statistical precision and computational feasibility. Initially, larger thresholds were used during preliminary analysis, after which we relaxed these thresholds to add experiments until the computational time required to extract the raw data became the binding constraint.

In November 2020, Meta (then Facebook) disclosed a bug affecting conversion metrics for about a year (Schiff, 2020). Our data were unaffected because we obtained them from an upstream source in the conversion measurement pipeline.

Figure A-1 presents the empirical distribution of several experiment characteristics. The figure illustrates that we observe significant variation in the mix of experiments in our dataset. Table A-2 displays the distribution of RCTs by the industry vertical of the advertiser.

Figure A-1: Distribution of Experiment Characteristics




Histogram excludes the top 1% of experiment population size; Dashed line shows median

Table A-2: RCT Counts by Vertical

| Vertical | N |
|---|---|
| Ecomm | 616 |
| Retail | 504 |
| Travel | 350 |
| EntMedia | 196 |
| CPG | 121 |
| Telecom | 105 |
| Tech | 102 |
| FinServ | 83 |
| ProfServ | 65 |
| Other | 36 |
| Automotive | 34 |
| Gaming | 14 |

# Appendix 3 Simulation Details

This appendix describes the simulation methodology used to evaluate PIE's predictive performance. Each simulation constructs potential outcomes for a set of campaigns $r \in \{1, \ldots, R\}$, each containing $N_r$ users. We use a linear probability model to generate user-level behaviors, ensuring that

treatment effects are invariant to baseline conversion heterogeneity.

### 3.1 Campaign-Level Parameter Generation

For each campaign $r$, we draw the following parameters that govern user behavior and outcomes:

**Exposure probability.** The probability of exposure $p_E^r$ is drawn from $p_E^r \sim U[p_E^{\text{low}}, p_E^{\text{high}}]$.

**Baseline conversion probability and treatment effect.** We generate correlated baseline conversion probability $\alpha_r$ and treatment effect $\tau_r$ (measured in percentage points) using a Gaussian copula. This method allows us to independently control (1) the marginal distribution of each variable and (2) the correlation between them.

The procedure works as follows. First, we draw $(Z_1, Z_2)$ from a bivariate normal distribution with correlation $\rho_{\alpha\tau}$:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{\alpha\tau} \\ \rho_{\alpha\tau} & 1 \end{pmatrix} \right) \tag{1}$$

This gives us two correlated normal random variables. Second, we apply the probability integral transformation to convert each normal variable to a uniform variable on $[0, 1]$:

$$U_1 = \Phi(Z_1), \quad U_2 = \Phi(Z_2) \tag{2}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Because $\Phi(\cdot)$ is a monotonic transformation, $U_1$ and $U_2$ remain correlated with approximately the same rank correlation as $Z_1$ and $Z_2$. Moreover, by the probability integral transformation theorem, both $U_1$ and $U_2$ are uniformly distributed on $[0, 1]$.

Finally, we transform these uniform variables to our desired marginal distributions using inverse transform sampling:

$$\alpha_r = \alpha^{\text{low}} + (\alpha^{\text{high}} - \alpha^{\text{low}}) \cdot U_1 \tag{3}$$

$$\tau_r = \tau^{\text{low}} + (\tau^{\text{high}} - \tau^{\text{low}}) \cdot U_2 \tag{4}$$

This maps each uniform variable to a uniform distribution on the specified range. These transformations preserve the dependence structure: $\alpha_r$ and $\tau_r$ inherit the correlation from $Z_1$ and $Z_2$, while having the desired marginal distributions (both uniform, but with different ranges).

This approach gives us precise control over both the marginal distributions (the ranges from which $\alpha_r$ and $\tau_r$ are drawn) and their correlation, which would be difficult to achieve with simple bivariate sampling methods.

**Effect of user propensity on outcomes.** The parameter $\gamma_r$ controls how strongly unobservable user propensity $U_i^r$ affects baseline conversion probability. We draw $\gamma_r \sim U[\gamma^{\text{low}}, \gamma^{\text{high}}]$.

**Effect of clicks on conversions.** The parameter $\delta_r$ captures the causal effect of clicking on conversion probability (in percentage points). We draw $\delta_r \sim U[\delta^{\text{low}}, \delta^{\text{high}}]$.

**Click probability.** The base click probability is drawn as $CP_r \sim U[CP^{\text{low}}, CP^{\text{high}}]$.

### 3.2 Individual-Level Behavior Generation

For each user $i$ in campaign $r$, we simulate the following behaviors:

**User propensity.** We draw unobservable user propensity $U_i^r \sim \text{Uniform}(0, 1)$.

**Exposure.** Following equation (10), user $i$ is exposed if and only if $U_i^r \geq 1 - p_E^r$:

$$D_i^r = \mathbb{1}\{U_i^r \geq 1 - p_E^r\}$$

**Click behavior.** The probability that user $i$ clicks (conditional on exposure) depends on user propensity:

$$\Pr(C_i^r = 1 | D_i^r = 1) = CP_r + \beta_{CP} \cdot U_i^r$$

where $\beta_{CP}$ is a fixed slope parameter we specify in our simulations. Users click only if exposed: $C_i^r = \mathbb{1}\{\text{click}\} \cdot D_i^r$.

**Potential outcomes.** Using a linear probability model, we generate outcomes for the treatment and control conditions. The underlying probabilities are:

$$\begin{aligned}
\text{Prob}_i^{r,T} &= \alpha_r + \gamma_r \cdot U_i^r + \tau_r \cdot D_i^r + \delta_r \cdot C_i^r \\
\text{Prob}_i^{r,C} &= \alpha_r + \gamma_r \cdot U_i^r
\end{aligned}$$

We set our simulation parameters such that these probabilities are bounded within $[0, 1]$. Finally, we draw binary outcomes:

$$\begin{aligned}
Y_{ti}^r &\sim \text{Bernoulli}(\text{Prob}_i^{r,T}) \\
Y_{ci}^r &\sim \text{Bernoulli}(\text{Prob}_i^{r,C})
\end{aligned}$$

**Last-click conversions.** We define a last-click conversion as $LCC_i^r = Y_{ti}^r \cdot C_i^r$, indicating that user $i$ both clicked and converted in the treatment condition.

### 3.3 Aggregation and Sample Construction

We aggregate individual-level data at the campaign level, yielding for each campaign $r$ the number of conversions in each group ($Y_{tr}, Y_{cr}$), clicks ($C_{tr} \geq 0$, $C_{cr} = 0$), last-click conversions ($LCC_{tr} \geq 0$, $LCC_{cr} = 0$), and the ad-exposure rate ($\phi_{tr} \geq 0$, $\phi_{cr} = 0$).

**Sample splitting.** To avoid overfitting in small samples, we implement sample splitting within each campaign.[26] For each campaign $r$, we randomly assign the $N_r$ users to two equally-sized samples. Sample 1 is used to calculate the observed incremental conversions $Y_{tr} - Y_{cr}$, which we use

[26] We thank Ryan Shyu for suggesting this.

to calculate $\hat{\psi}_r$, the ATTs that are the dependent variable in the PIE regression. Sample 2 is used to calculate the predictors ($Y_{tr}$, $\phi_{tr}$, $C_{tr}$, $LCC_{tr}$). This ensures that predictors and outcomes are calculated from independent samples, eliminating mechanical correlation due to sampling error.

**Train-test split.** After generating data for all campaigns and calculating both outcomes and predictors, we split campaigns into equally-sized training and test sets. We estimate PIE specifications on the training sample using OLS regression:

$$\hat{\psi}_r = \theta_0 + \theta_1 \cdot Y_{tr} + \theta_2 \cdot \phi_{tr} + \theta_3 \cdot C_{tr} + \theta_4 \cdot LCC_{tr} + \nu_r \tag{5}$$

where all right-hand side variables are observable from the treatment group alone. Using estimated coefficients from the training sample, we generate predictions for the test sample:

$$\hat{\psi}_r^{\text{PIE}} = \hat{\theta}_0 + \hat{\theta}_1 \cdot Y_{tr} + \hat{\theta}_2 \cdot \phi_{tr} + \hat{\theta}_3 \cdot C_{tr} + \hat{\theta}_4 \cdot LCC_{tr} \tag{6}$$

### 3.4 Evaluation Metrics

We evaluate prediction accuracy using $R^2$, which measures the proportional reduction in mean squared error relative to a naive baseline that predicts the mean treatment effect for every campaign:

$$R^2 = 1 - \frac{\sum_r (\hat{\psi}_r - \hat{\psi}_r^{\text{PIE}})^2}{\sum_r (\hat{\psi}_r - \bar{\psi})^2} \tag{7}$$

where $\bar{\psi}$ is the mean ATT across test campaigns. Higher $R^2$ indicates that the model substantially outperforms the naive baseline; $R^2 = 0$ indicates no improvement over predicting the mean.

We report $R^2$ for models with increasingly rich feature sets: total conversions only ($R^2_{y_t}$), adding exposure fraction ($R^2_{y_t,\phi}$), adding clicks ($R^2_{y_t,\phi,C}$), and the full model including last-click conversions ($R^2_{y_t,\phi,C,LCC}$). Each simulation is repeated 100 times with different random seeds, and we pool all test observations across repetitions to calculate a single $R^2$ for each scenario.

### 3.5 Detailed Simulation Results

**Number and Size of RCTs** Table A-3 shows how predictive performance improves with the number of RCTs available for training. With 50 RCTs, the full model achieves $R^2 = 0.260$, improving to 0.442 with 1,000 RCTs and 0.448 with 10,000 RCTs. This occurs because larger training samples allow the model to better distinguish signal from noise. The gains diminish at larger sample sizes, with the increment from 1,000 to 10,000 RCTs adding little to $R^2$. Note that the number of RCTs at which gains diminish is highly dependent on other parameters of the simulation. For example, the larger the treatment effect, the fewer RCTs are needed.

The second panel examines RCT size, holding the number of RCTs fixed at 1,000. Increasing the sample size per RCT from 1,000 to 10,000 users improves $R^2$ from 0.190 to 0.442. This occurs because larger RCTs produce more precise estimates of $\psi_r$ and the predictor variables in the PIE models.

**Heterogeneity in Baseline Conversion Rates and Treatment Effects** Table A-4 examines how heterogeneity in $\alpha_r$ and $\tau_r$ affects the performance of PIE. Panel A varies baseline conversion

Table A-3: PIE Performance: Varying Number and Size of RCTs

| | $R^2$ for PIE model using | | | |
|---|---|---|---|---|
| Scenario | $y_t$ | $y_t, \phi$ | $y_t, \phi, C$ | $y_t, \phi, C, LCC$ |
| *Panel A: Varying Number of RCTs (size = 10,000)* | | | | |
| 50 RCTs | 0.385 | 0.385 | 0.331 | 0.260 |
| 100 RCTs | 0.412 | 0.412 | 0.396 | 0.368 |
| 1,000 RCTs | 0.446 | 0.446 | 0.444 | 0.442 |
| 10,000 RCTs | 0.448 | 0.448 | 0.448 | 0.448 |
| *Panel B: Varying Size of RCTs (N = 1,000)* | | | | |
| 1,000 users | 0.213 | 0.213 | 0.201 | 0.190 |
| 5,000 users | 0.400 | 0.400 | 0.396 | 0.392 |
| 10,000 users | 0.446 | 0.446 | 0.444 | 0.442 |

*Notes:* Out-of-sample $R^2$ for PIE models. Panel A varies the number of RCTs while holding RCT size constant at 10,000 users. Panel B varies RCT size while holding the number of RCTs constant at 1,000. All scenarios assume full compliance ($p_E = 1.00$), $\alpha_r \sim U[0.13, 0.17]$, and $\tau_r \sim U[0.02, 0.06]$.

probability ($\alpha$) heterogeneity while holding $\tau_r \sim U[0.02, 0.06]$. The full model's $R^2$ declines from 0.666 (low heterogeneity) to 0.282 (high). This occurs because greater heterogeneity in $\alpha_r$ increases the variance of $Y_{tr}$ without proportionally increasing our ability to predict $\psi_r$, making the naive mean-based baseline harder to beat, consistent with what we found in Section 3.4.

Panel B varies treatment effect ($\tau_r$) heterogeneity with $\alpha_r \sim U[0.13, 0.17]$. The full model's $R^2$ increases from 0.152 (low heterogeneity) to 0.640 (high). When treatment effects vary substantially, the error of the naive baseline (mean prediction) increases, while the PIE model's predictions exploit the fact that $\tau$ is on both sides of equations 17 and 19.

Table A-4: PIE Performance: Varying Heterogeneity in $\alpha_r$ and $\tau_r$

| | $R^2$ for PIE model using | | | |
|---|---|---|---|---|
| Scenario | $y_t$ | $y_t, \phi$ | $y_t, \phi, C$ | $y_t, \phi, C, LCC$ |
| *Panel A: Varying Heterogeneity of Baseline Conversion Rate ($\alpha_r$)* | | | | |
| Low: $U[0.14, 0.16]$ | 0.669 | 0.669 | 0.668 | 0.666 |
| Moderate: $U[0.13, 0.17]$ | 0.449 | 0.449 | 0.447 | 0.445 |
| High: $U[0.12, 0.18]$ | 0.286 | 0.286 | 0.284 | 0.282 |
| *Panel B: Varying Heterogeneity of Treatment Effects ($\tau_r$)* | | | | |
| Low: $U[0.03, 0.05]$ | 0.161 | 0.161 | 0.156 | 0.152 |
| Moderate: $U[0.02, 0.06]$ | 0.440 | 0.440 | 0.439 | 0.437 |
| High: $U[0.01, 0.07]$ | 0.643 | 0.643 | 0.642 | 0.640 |

*Notes:* Out-of-sample $R^2$ for PIE models under varying degrees of parameter heterogeneity. Panel A varies baseline conversion probability range while holding $\tau_r \sim U[0.02, 0.06]$ constant. Panel B varies treatment effect heterogeneity while holding $\alpha_r \sim U[0.13, 0.17]$ constant. All scenarios use 1,000 RCTs with 10,000 users each and full compliance ($p_E = 1.00$).

**Correlation Between $\alpha_r$ and $\tau_r$** Table A-5 varies the correlation from $-0.9$ to $+0.9$ while holding the marginal distributions at $\alpha_r \sim U[0.13, 0.17]$ and $\tau_r \sim U[0.02, 0.06]$. Positive correlation improves $R^2$ substantially, increasing from 0.440 (zero correlation) to 0.886 (correlation = 0.90). This occurs because positive correlation amplifies the systematic signal in $Y_{tr}$: when campaigns with high baseline conversion rates also have large treatment effects, $Y_{tr}$ provides a strong signal about both components.

Conversely, negative correlation severely degrades performance, with $R^2$ falling to 0.016 at correlation $= -0.90$. Strong negative correlation means that in $Y_{tr} = \alpha_r + \tau_r + \varepsilon_{tr}$, $\alpha_r$ and $\tau_r$ nearly cancel each other out, leaving only noise as an explanatory variable.

Table A-5: PIE Performance: Varying Correlation Between $\alpha_r$ and $\tau_r$

| | $R^2$ for PIE model using | | | |
|---|---|---|---|---|
| Correlation | $y_t$ | $y_t, \phi$ | $y_t, \phi, C$ | $y_t, \phi, C, LCC$ |
| *Panel A: Positive Correlation* | | | | |
| $\rho = 0.00$ | 0.443 | 0.443 | 0.442 | 0.440 |
| $\rho = 0.45$ | 0.660 | 0.660 | 0.659 | 0.658 |
| $\rho = 0.90$ | 0.888 | 0.888 | 0.887 | 0.886 |
| *Panel B: Negative Correlation* | | | | |
| $\rho = 0.00$ | 0.447 | 0.447 | 0.445 | 0.443 |
| $\rho = -0.45$ | 0.234 | 0.234 | 0.230 | 0.228 |
| $\rho = -0.90$ | 0.021 | 0.021 | 0.019 | 0.016 |

*Notes:* Out-of-sample $R^2$ for PIE models under varying correlation between baseline conversion rates and treatment effects. Marginal distributions held constant at $\alpha_r \sim U[0.13, 0.17]$ and $\tau_r \sim U[0.02, 0.06]$. Correlation induced using a Gaussian copula. All scenarios use 1,000 RCTs with 10,000 users each and full compliance ($p_E = 1.00$).

**Effect of Clicks on Conversions ($\delta_r$)** Table A-6 examines scenarios where clicking has a direct causal effect on conversion probability, parameterized by $\delta_r$.

Under low click probability ($CP_r \sim U[0, 0.1]$, $\beta_{CP} = 0.1$), introducing $\delta_r$ heterogeneity has minimal impact, with $R^2$ remaining around 0.45–0.46 across all scenarios. When few users click, these features contain little information regardless of click effect variation.

Under high click probability ($CP_r \sim U[0, 0.3]$, $\beta_{CP} = 0.5$), $R^2$ increases from 0.513 to 0.657 as $\delta_r$ heterogeneity increases. The incremental contribution of last-click conversions grows substantially with $\delta$ heterogeneity, adding only 0.001 under low heterogeneity but 0.057 under high heterogeneity.

**Selective Exposure and Unobserved Heterogeneity** We now examine how PIE performance depends on unobserved heterogeneity in baseline outcomes and selection into exposure. Recall from equation (10) that exposure is determined by $D_i^r = \mathbb{1}\{U_i^r \geq 1 - p_r(Z_i^r)\}$, where $U_i^r \sim \text{Uniform}(0, 1)$ governs selection into exposure. The parameter $\gamma_r$ controls how strongly $U_i^r$ affects baseline conversion probability. When $\gamma_r = 0$, user propensity determines exposure but has no effect on outcomes. When $\gamma_r > 0$, the same unobservable that determines exposure also affects baseline conversion rates.

Table A-6: PIE Performance: Varying Click Effects on Conversions ($\delta_r$)

| | $R^2$ for PIE model using | | | |
|---|---|---|---|---|
| Scenario | $y_t$ | $y_t, \phi$ | $y_t, \phi, C$ | $y_t, \phi, C, LCC$ |
| *Panel A: Low Click Probability,* $CP_r \sim U[0.00, 0.10]$, $\beta_{CP} = 0.10$ | | | | |
| Low Het: $\delta_r \sim U[0.04, 0.06]$ | 0.448 | 0.448 | 0.451 | 0.448 |
| Mod Het: $\delta_r \sim U[0.03, 0.07]$ | 0.453 | 0.453 | 0.455 | 0.453 |
| High Het: $\delta_r \sim U[0.01, 0.09]$ | 0.458 | 0.458 | 0.460 | 0.464 |
| *Panel B: High Click Probability,* $CP_r \sim U[0.00, 0.30]$, $\beta_{CP} = 0.50$ | | | | |
| Low Het: $\delta_r \sim U[0.04, 0.06]$ | 0.483 | 0.483 | 0.514 | 0.513 |
| Mod Het: $\delta_r \sim U[0.03, 0.07]$ | 0.511 | 0.511 | 0.537 | 0.548 |
| High Het: $\delta_r \sim U[0.01, 0.09]$ | 0.588 | 0.588 | 0.600 | 0.657 |

*Notes:* Out-of-sample $R^2$ for PIE models when clicks have a causal effect on conversions. Panel A examines low baseline click probability; Panel B examines high click probability. Within each panel, we vary $\delta_r$ heterogeneity. Click probability (conditional on exposure) depends on user propensity: $\Pr(C_i^r = 1 | D_i^r = 1) = CP_r + \beta_{CP} \cdot U_i^r$. All scenarios use 1,000 RCTs with 10,000 users each, $\alpha_r \sim U[0.13, 0.17]$, $\tau_r \sim U[0.02, 0.06]$, and full compliance ($p_E = 1.00$).

i. **Effect of $\gamma_r$ Heterogeneity.**

Panel A in Table A-7 shows results with partial selection into exposure ($p_E \sim U[0.4, 0.6]$). The full model's $R^2$ declines from 0.568 (low $\gamma_r$ heterogeneity) to 0.353 (high heterogeneity). This is analogous to increasing heterogeneity in the baseline probability $\alpha_r$: greater heterogeneity in $\gamma_r$ increases the variance of $Y_{tr}$ without proportionally increasing our ability to predict $\psi_r$ because $\gamma_r$ is only on the right-hand side of equation 17.

ii. **Effect of Exposure Selection Under Different $\gamma_r$ Scenarios.**

Panel B in Table A-7 varies exposure from high ($p_E \sim U[0.7, 0.9]$) to highly selective ($p_E \sim U[0.1, 0.3]$) for the moderate $\gamma_r \sim U[0.10, 0.20]$ case. The full model's $R^2$ increases from 0.319 for high exposure probability, to 0.430 for medium exposure probability, and 0.737 for low exposure probability.

This substantial improvement occurs because as selection increases, only users with increasingly high $U_i^r$ are exposed, and $\phi_r$ becomes increasingly informative about user composition, which PIE can exploit relative to a prediction of the mean. The table shows that PIE's use of the exposure fraction ($\phi_r$) as a post-determined feature is helpful when the selection mechanism correlates with outcomes. When $\gamma_r > 0$, the unobservable that determines exposure also affects baseline outcomes, and $\phi_r$ becomes increasingly valuable as a proxy for user composition.

iii. **Click Features as Proxies for User Composition.**

Panel C demonstrates an important property of click-based features: they improve predictions even when clicks have no direct causal effect on conversions ($\delta_r = 0$). Under moderate unobserved heterogeneity ($\gamma_r \sim U[0.10, 0.20]$) and partial exposure ($p_E \sim U[0.4, 0.6]$), adding click features increases $R^2$ from 0.087 (using only $y_t$) to 0.196 (adding $\phi$), and further to 0.252 (adding clicks and last-click conversions).

This occurs because under partial exposure, clicks reveal information about user composition. Users with high $U_i^r$ are both more likely to be exposed and, through the $\gamma_r$ parameter, more likely to convert. Since click probability also depends on $U_i^r$ (via $\Pr(C_i^r = 1|D_i^r = 1) = CP_r + \beta_{CP} \cdot U_i^r$), click features serve as a proxy for the unobserved user characteristics that drive both exposure and outcomes. This demonstrates that click features can be valuable predictors of treatment effects even in the absence of any causal click-to-conversion pathway.

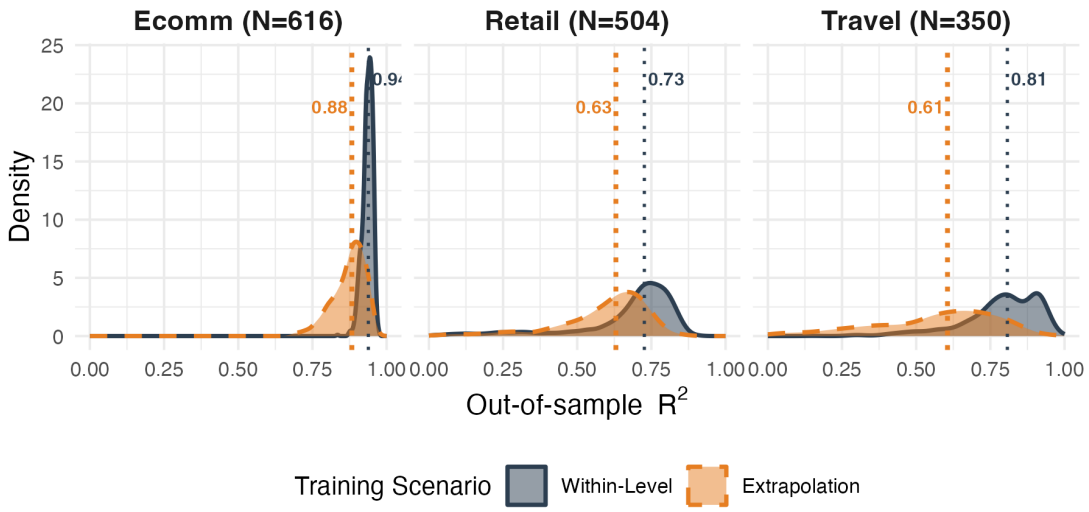
Table A-7: PIE Performance: Heterogeneity $\gamma_r$ and Exposure Selection

| | $R^2$ for PIE model using | | | |
|---|---|---|---|---|
| Exposure Range | $y_t$ | $y_t, \phi$ | $y_t, \phi, C$ | $y_t, \phi, C, LCC$ |
| *Panel A: Varying Heterogeneity in the effect of $U_i$* | | | | |
| Low Heterogeneity: $\gamma_r \sim U[0.14, 0.16]$ | 0.244 | 0.357 | 0.405 | 0.568 |
| Moderate Heterogeneity: $\gamma_r \sim U[0.10, 0.20]$ | 0.126 | 0.276 | 0.337 | 0.460 |
| High Heterogeneity: $\gamma_r \sim U[0.00, 0.30]$ | 0.022 | 0.210 | 0.285 | 0.353 |
| *Panel B: Varying Exposure Selection* | | | | |
| $p_E \sim U[0.70, 0.90]$ | 0.226 | 0.272 | 0.276 | 0.319 |
| $p_E \sim U[0.40, 0.60]$ | 0.120 | 0.284 | 0.289 | 0.430 |
| $p_E \sim U[0.10, 0.30]$ | 0.051 | 0.655 | 0.656 | 0.737 |
| *Panel C: Click Features Without Click Treatment Effect ($\delta_r = 0$)* | | | | |
| $\gamma_r \sim U[0.10, 0.20]$, $p_E \sim U[0.40, 0.60]$ | 0.087 | 0.196 | 0.193 | 0.252 |

*Notes:* Out-of-sample $R^2$ under varying degrees of exposure selection and unobserved heterogeneity. Panels A and B use $\delta_r \sim U[0.03, 0.07]$. Panel C demonstrates that click features ($C$ and $LCC$) improve predictions even when clicks have no causal effect on conversions ($\delta_r = 0$), because they proxy for user composition under partial exposure. All scenarios use 1,000 RCTs with 10,000 users each, $\alpha_r \sim U[0.13, 0.17]$, and $\tau_r \sim U[0.02, 0.06]$.

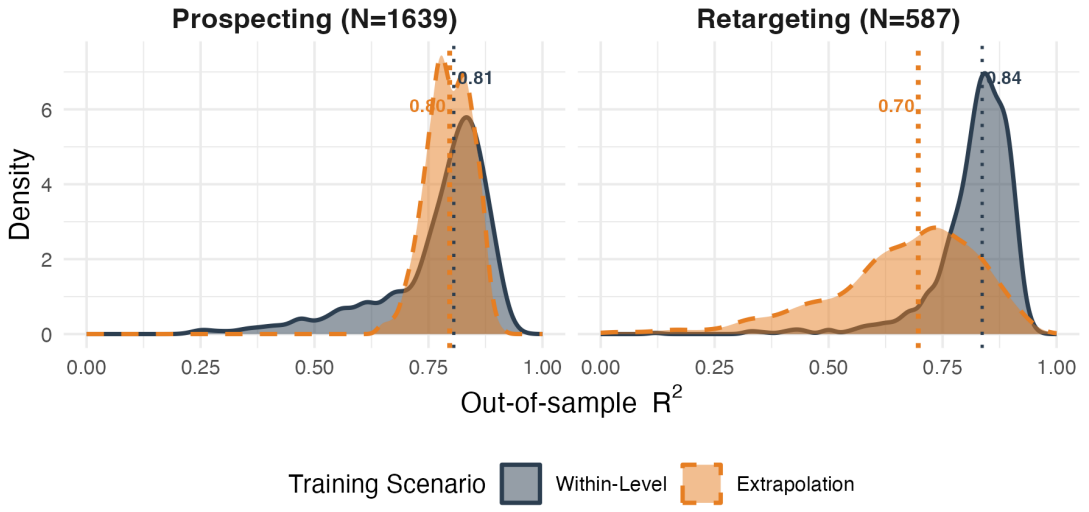
# Appendix 4 Additional Extrapolation by Segment Figures

Figure A-2: Extrapolation performance by advertiser vertical



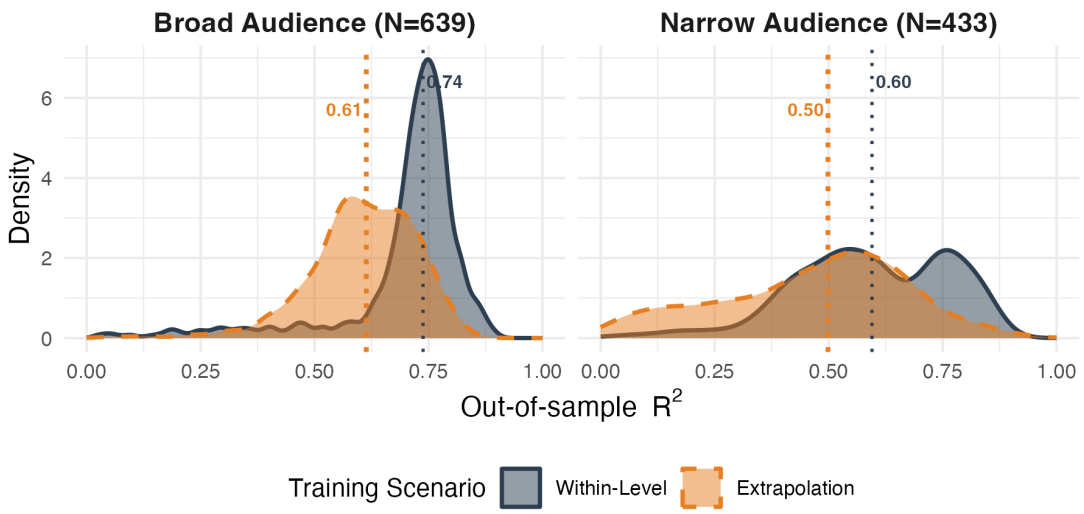

*Notes:* Empirical distributions of out-of-sample $R^2$ from the segment-level hold-out experiment, shown here for a subset of advertiser verticals. Each panel corresponds to a vertical (e-commerce, retail, travel), with separate density curves for models trained within the focal group (red) and for models trained by extrapolating from the remaining groups (blue). The distributions are based on 1,000 repeated random partitions of the data, providing a view of how predictive accuracy varies across resamples under the within- and extrapolation-training conditions. Vertical dotted lines indicate the median $R^2$ within each training scenario.

Figure A-3: Extrapolation performance for prospecting vs. retargeting campaigns



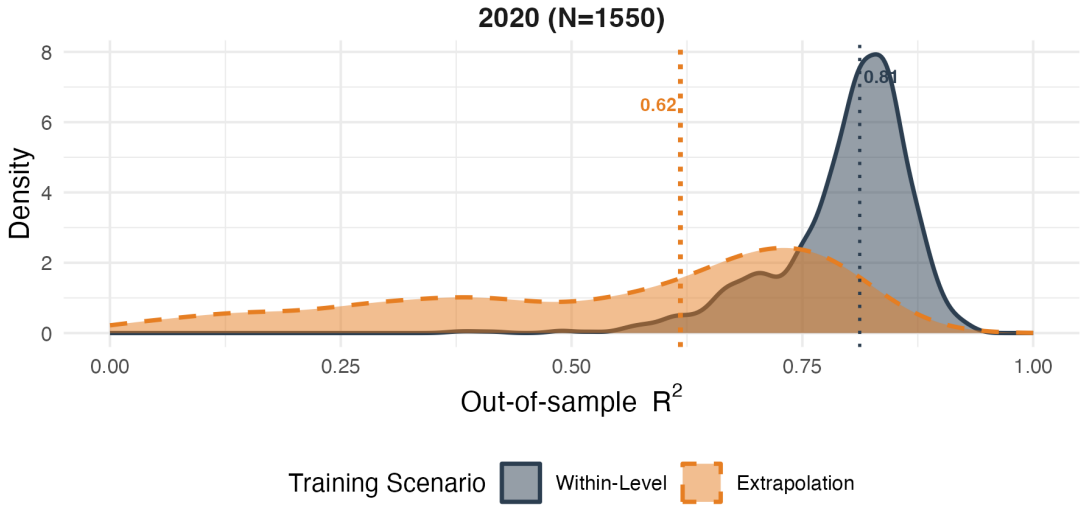

*Notes:* Empirical distributions of out-of-sample $R^2$ from the segment-level hold-out experiment, shown here based on whether the campaign is focused on reaching new potential customers (**Prospecting**) or aims to reach existing customers (**Retargeting**). In each panel, there are separate density curves for models trained within the focal group (red) and for models trained by extrapolating from the remaining groups (blue). The distributions are based on 1,000 repeated random partitions of the data, providing a view of how predictive accuracy varies across resamples under the within- and extrapolation-training conditions. Vertical dotted lines indicate the median $R^2$ within each training scenario.

Figure A-4: Extrapolation performance for custom audience campaigns



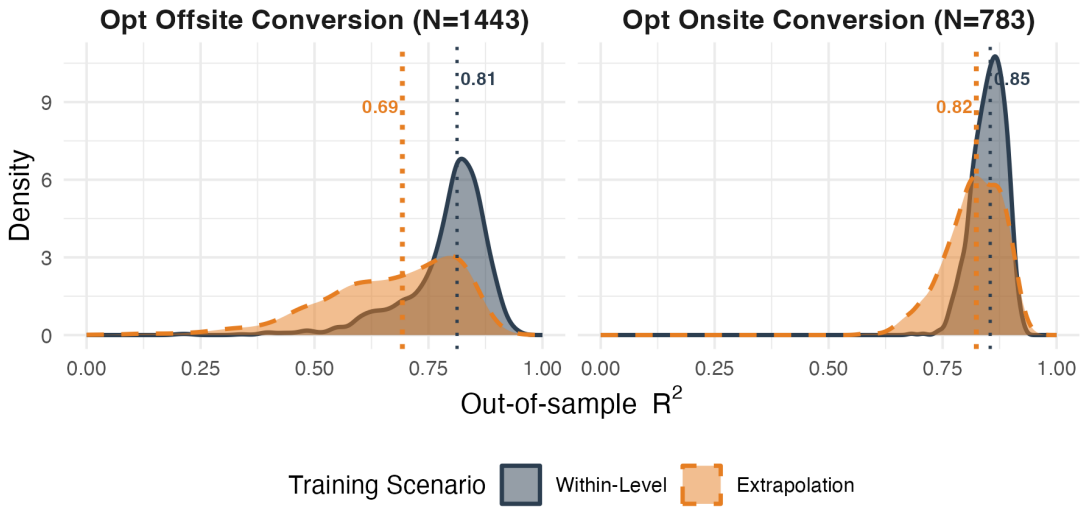

*Notes:* Empirical distributions of out-of-sample $R^2$ from the segment-level hold-out experiment, shown here based on whether the campaign targets a **narrow** custom audience or a **broad** custom audience. In each panel, there are separate density curves for models trained within the focal group (red) and for models trained by extrapolating from the remaining groups (blue). The distributions are based on 1,000 repeated random partitions of the data, providing a view of how predictive accuracy varies across resamples under the within- and extrapolation-training conditions. Vertical dotted lines indicate the median $R^2$ within each training scenario.

Figure A-5: Extrapolation performance training on campaigns in 2019 to extrapolate to campaigns run in 2020




*Notes:* Empirical distributions of out-of-sample $R^2$ from the segment-level hold-out experiment. In this case, the within (red) trains a model on campaigns from 2020 and predicts on campaigns from 2020, whereas the extrapolation (blue) trains the model on campaigns from 2019 and predicts on campaigns from 2020. The distributions are based on 1,000 repeated random partitions of the data, providing a view of how predictive accuracy varies across resamples under the within- and extrapolation-training conditions. Vertical dotted lines indicate the median $R^2$ within each training scenario.

Figure A-6: Extrapolation performance by type of conversion optimization (onsite vs. offsite)




*Notes:* Empirical distributions of out-of-sample $R^2$ from the segment-level hold-out experiment, shown here based on whether the campaign is optimizing for offsite conversions or onsite conversions. In each panel, there are separate density curves for models trained within the focal group (red) and for models trained by extrapolating from the remaining groups (blue). The distributions are based on 1,000 repeated random partitions of the data, providing a view of how predictive accuracy varies across resamples under the within- and extrapolation-training conditions. Vertical dotted lines indicate the median $R^2$ within each training scenario.