The A/B Test Deception: Divergent Delivery, Ad Response Heterogeneity, and Erroneous Inferences in Online Advertising Field Experiments

Michael Braun
Cox School of Business
Southern Methodist University
braunm@smu.edu

Eric M. Schwartz
Ross School of Business
University of Michigan
ericmsch@umich.edu

July 28, 2021

Abstract

Advertisers and researchers use tools provided by advertising platforms to conduct randomized experiments for testing user responses to creative elements in online ads. Internally valid comparisons between ads require the mix of experimental users exposed to each ad to be similar across all ads. But that internal validity is threatened when platforms' targeting algorithms deliver each ad to its own optimized mix of users, which diverges across ads. We extend the potential outcomes model of causal inference to treat random assignment of ads and the user exposure states for each ad as two separate decisions. We then demonstrate how targeting ads to users leads advertisers to incorrectly infer which ad performs better, based on aggregate test results. Through analysis and simulation, we characterize how bias in the aggregate estimate of the difference between two ads' lifts is driven by the interplay between heterogeneous responses to different ads and how platforms deliver ads to divergent subsets of users. We also identify conditions for an undetectable "Simpson's reversal," in which all unobserved types of users may prefer ad A over ad B, but the advertiser mistakenly infers from aggregate experimental results that users prefer ad B over ad A.

Keywords: Targeted online advertising, A/B testing, measuring advertising effectiveness, causal inference, experimental design, Simpson's paradox, social media

1 Introduction

Many online publishers provide tools to help advertisers conduct randomized advertising experiments on the publisher's own ad platform. These ad experimentation tools ostensibly enable advertisers to learn about their ads' effects on users in the publisher's environment. But these tools also benefit the publisher by providing evidence for its two key value propositions to the advertiser: *delivering* ads to users, and *targeting* the "right" ads to the "right" users.

One reason advertisers run such experiments is to learn about how users respond to different *creative elements* of ads (e.g., copy, images, and message). The purpose of their experiments is *inference*. These advertisers want to compare performance across multiple ads in a live campaign by isolating those ad effects from the impact of how the targeting algorithm matches ads to specific types of users. This applies across a variety of "advertiser" settings. For example, academic consumer behavior researchers test hypotheses about psychological constructs which are operationalized as creative elements in each ad treatment. Commercial advertisers may use online ad experiments to learn which creative elements of ads — or even inputs used in generating those creatives (e.g., ad agencies) — yield the best response among a particular segment of users, and then they may apply those inferences to ads on other platforms, on other media, or elsewhere in the marketing mix. Policy makers also want to know how an ad algorithm treats users differently, especially when testing ads for public services, like housing or employment. This motivation for online experiments is driven by questions about learning which creative elements best explain *why* users may be more likely to respond to some ads than others.

But some advertisers' motive for running experiments is not inference, but rather *prediction*. The experiment answers the question "which ad will be best?" just to deliver that better performing ad after the test. The advertiser with this "test and roll" objective (as in Feit and Berman 2019) may not care why ad A generates more conversions than ad B, whether because ad A contains more persuasive and compelling creative elements than B, or because the platform serves ad A to a targeted mix of users who are more amenable to responding to that ad than the mix who are exposed to B are responsive to their ad. That advertiser only cares about exposing users on the platform to ad A because it is "better."

This paper's aim is to provide a caution about available state-of-the-art available ad experimentation tools to advertisers whose initial goal is inferential. The problem with inferences drawn from online ad experiments, as if they were more controlled experiments, is the following. In a targeted ad environment, the types of users exposed to each of the experiment's ad treatments may be different for each ad. Thus, the observed results (e.g., incremental lift observed among users exposed to each ad) contain a confound between each ad's true ability to generate user response, and the platform's targeting algorithm. As a result, a so-called A/B/n test designed to compare effects across multiple ads is no longer a straightforward randomized controlled trial when conducted on a targeted ad platform. In targeted ad settings, the common A/B/n test is neither a direct comparison of

creatives "A,B,...,n" nor a randomized "test" measuring those creatives' effects. If the algorithm targets ad *A* to one user type expected respond better to *A*, and ad *B* to another user type expected to respond better to *B*, then assigned exposure to experimental ad treatments is no longer random, and comparisons between ads suffer from threats to internal validity.

This feature of targeting different ads to different users has been called "divergent delivery" (Johnson 2020) and "skewed delivery" (Ali et al. 2019). We will use the first of these terms, and define *divergent delivery* as the degree to which one ad's mix of targeted users differs from other ads' mixes of targeted users. This concept and its effects on inference from online experiments are not new, as they have been discussed in context of online experimentation by, among others, Eckles et al. (2018), Ali et al. (2019), and Johnson (2020). But the interplay between divergent delivery and the user-ad response interaction has not been studied in a formal sense. Our contributions to the literature arise as we answer the following four questions.

How can we formally define divergent delivery? In Sec. 2, we introduce a mathematical framework that characterizes how a targeting algorithm engaging in divergent delivery creates an interaction between the effects of two separate decisions about targeted ad delivery: which users to favor and which ads to favor. Our framework builds on the Rubin potential outcomes model for casual inference (RCM, Rubin 1974). But this prominent mathematical framework for causal inference and its many extensions do not accommodate some of the practical realities of doing online ad experiments in the field. Recent research uses the Rubin potential outcomes model to communicate the value of randomized experiments over observational methods to measure the effect of exposure to an ad in a targeted ad environment (Gordon et al. 2019). While the treatment-vs-control design in that paper appropriately measures the impact of a *single* ad on a population targeted with that ad (i.e., intended-to-be-treated users), the mathematical framework is not suited to accommodate comparisons *across* ads delivered to groups targeted differently. Our proposed framework extends the RCM to accommodate all the following: (1) multiple (two or more) ad treatments; (2) potential outcomes for every combination of ad treatment (*A*, *B*, etc.) and exposure status (exposed/unexposed); (3) different levels of assignment non-compliance (user and algorithmic); (4) two levels of user randomization (initial ad treatment and exposure conditional on targeting), (5) users' heterogeneous responses to those ads; and (6) platform's targeting of those ads with divergent delivery.

When is divergent delivery a problem? Divergent delivery creates a problem when the advertiser-experimenter wants to infer a causal comparison of the effects of two or more ads. When an experiment is designed to measure the effect of an ad by comparing responses to the focal ad with responses to a control (placebo) ad (as we will show as Design 1 in Sec. 2.2), the targeting algorithm will cause the mixes of users exposed to the treatment and control ads will differ (Gordon et al. 2019). Researchers in academia and industry have proposed *and implemented* tools that provide a remedy (Design 2 in Sec. 2.2). The "ghost ads" approach, proposed by Johnson et al. (2017a) and implemented at Google, randomly splits users targeted with the focal ad into a group to be exposed to the focal ad,

and a group who will not see that ad, but instead see the next-best ad in the auction. Both groups' behaviors are tracked and reported, allowing for inferences about the incremental outcomes by comparing otherwise identical exposed and unexposed groups. Similarly, Facebook provides tools that let advertisers run "holdout tests" in which exposed and unexposed users are randomly sampled from the same population of targeted users (Gordon et al. 2019).

Ghost ads and holdout tests are appealing for measuring within-ad effects (i.e., "this ad" vs "not this ad"), but even the above-cited authors recognize that these test designs do not resolve the targeting problem when *comparing effects between ads*. This problem is similar to the more familiar issue of covariate imbalance across treatments when user traits can be *observed*. Ali et al. (2019), recognizing the issue of divergent delivery on observable characteristics, unveil how targeting changes the mix of users exposed to different ads for housing and employment opportunities, with important implications for public policy and equity.[1] When this confound occurs during the course of an experiment, the internal validity of the study is threatened. Eckles et al. (2018) comment on an experiment by Matz et al. (2017), delivered through Facebook's experimentation platform, which was designed to test consumer responses to psychological constructs operationalized with different ad copy. Although exposed users were implicitly assumed to be randomly assigned to ads, the distributions of *reported* covariates (e.g., gender, age) varied across ads (Eckles et al. 2018, Fig. 1). Therefore, inferences from the experimental data cannot speak to the researchers' question of interest, for the researchers' population of interest.

Yet, researchers continue to advocate for online experimentation tools as a way to test behavioral constructs in the field. For example, Orazi and Johnston (2020) demonstrate how to use Facebook's platform to run an A/B/n test that operationalizes consumer psychology theories about response to ads for COVID face masks. Kupor et al. (2015) and Kupor and Laurin (2020) manipulated creative elements of ads on Facebook, and Cecere et al. (2018) randomize users to ads using Snapchat. The distinguishing feature of these kinds of studies is that the advertiser (an academic researcher) collects data from an experiment on a publisher's platform, and infers causal comparisons about how an audience responds to characteristics of creative elements in ads to apply those insights to other settings (e.g., other platforms or advertising channels). Google, AdRoll, MediaMath, and Microsoft, among others, provide similar services (Gordon et al. 2021).

For such experiments, if the distribution of user types exposed to each ad were observable to the advertiser, then it might be possible for the advertiser to control for covariate imbalance with common observational studies methods (e.g., propensity score weighting). But in practice, the targeting algorithm relies primarily on *unobservable* information about users, and advertisers have no extant remedy. Since the platform's targeting algorithm is proprietary, the advertiser-experimenter will never know if or how much the mixes of targeted users exposed to each ad may differ due to unobservable divergent delivery. For our purposes, observed characteristics

[1] In a lawsuit, the US Department of Housing and Urban Development (HUD) claims Facebook's targeting algorithms excluded audiences for housing ads without HUD requesting it or knowing (Hao 2021).

either define the bounds of the *audience* (the subject pool for the experiment, such as "users in California who are interested in gardening") or comprise subsets of users in the results reported to the advertiser (coarse demographic groups such as age, gender, and location). We treat all other user traits that drive the likelihood a user is targeted with or responds to an ad as unobserved by the advertiser. The advertiser gets results that are aggregated across unobserved factors, even if the mixes of users targeted with each ad are different. What portion of the differences in results across ads are due to ad creatives alone or due to the way the targeting algorithm interacts with each ad creative differently? The advertiser will not know. It is impossible for the advertiser know how problematic divergent delivery could be for them, and it can lead them astray in their conclusions about why some ads perform better, and even which ad is better for any given user type.

The advertiser's perspective has not received as much attention as the platform's perspective in this literature. Our work is unique in that it reflects the advertiser's limited-information point of view by quantifying the consequences of platforms divergently delivering ads to heterogeneous users who differ in ways that are undetectable to the advertiser. This paper will demonstrate this thesis: *A/B/n tests designed to learn the relative effectiveness of creative elements of ads, as currently conducted in a targeted advertising environment, generate biased comparisons between ads.* In Sec. 3, we use our framework to mathematically describe outcomes, effects, targeting policies, and audience characteristics, building to a definition of a *bias* that captures the difference between the estimated and the true difference in effects of ads A and B on the advertiser's predefined audience. In Sec. 4, using both mathematical analysis and numerical simulation, we show how the direction and magnitude of the bias arise from the interplay between divergent delivery targeting policies and patterns of heterogeneity in users' response propensities. We establish the conditions that generate bias, including conditions for a *Simpson's reversal* in which ad A is more effective than ad B for all groups of users with similar types, but the effect an advertiser infers from aggregate experimental data (e.g., A better than B) *has the wrong sign*.[2] The simulation will show how (and when) divergent delivery and response heterogeneity collude to induce bias in A/B comparisons of lift among targeted users that mislead the advertiser in interpretation of their results.

What can be done, if anything, to mitigate targeting-induced biases? In Sec. 5, we analyze the implications of our alternative targeting policy that "shuts off" divergent delivery, and targets all ads in the experiment, as a unit, to a single mix of users. Through simulation, we compare how conversions and bias differ by policies with and without divergent delivery. By quantifying the economic gains and losses from disabling divergent delivery, we can consider the incentives for publishers and advertisers to keep it in place.

What are we not doing in this paper? Before continuing, we want to establish boundaries around the scope of the paper. We are not concerned with ad campaigns whose goals are anything other than direct response (e.g.,

[2]Blyth (1972) describe this pattern of aggregation bias as "Simpson's paradox," by which it is more commonly known. But a paradox that can be explained mathematically is "resolved," and thus is no longer paradoxical (Pearl 2014).

branding), nor are we addressing competitive or strategic considerations. This paper is not about inner-workings of targeting algorithms, adaptive experimental design, optimal bidding strategies, or other topics in the "ad tech stack." Also, the paper was not written with any one platform in mind. Instead, we remain tightly focused on a broadly applicable problem that arises from making inferences using experimental data from a platform that replicates the "production" case when targeting and divergent delivery are enabled.

2 Online ad experiments and causal inference

To begin, we define a set of terms, concepts, and experimental designs to characterize the targeted online advertising testing problem. Advertisers run *campaigns* on ad *platforms* owned by *publishers*. A campaign involves n_Z *ads*, labeled $Z \in \{z_1, z_2, \dots, z_{n_Z}\}$, where each ad is a bundle of creative elements, such as message, copy, and imagery.[3] When initializing a campaign, the advertiser defines an *audience* of users to whom the platform may deliver ads. [4] The user eventually generates an observed *outcome* $Y_i^{(\text{obs})}$, such as ad clicks, page views, or, as in our simulation in Sec. 4.2, a binary indicator of conversion. At the end of (or during) the campaign, the platform provides the advertiser a *report* that summarizes aggregated user outcomes, along with other data, such as the number of users exposed to each ad.

An *experiment* is a type of campaign that may include multiple ads, and lets the advertiser infer and compare how exposure to each ad affects $Y_i^{(\text{obs})}$. To run the experiment, the platform randomly assigns *every user in the audience* to exactly one ad treatment Z_i, with assignment probabilities $\boldsymbol{\zeta} = \{\zeta_{z_1}, \dots, \zeta_{z_{n_Z}}\}$, as the *ad treatment*.[5] This initial random assignment makes the user *eligible* to see the assigned ad, and only that ad, unlike in a non-experimental campaign where the user may see any or all of the n_Z ads. The user is *exposed* to the ad if the platform successfully delivers the ad impression.[6] Whether a user is exposed depends on many factors, including the experimental design, the platform's targeting algorithm, and the user's own behavior, all of which we will return to shortly. We indicate if a user is actually exposed to the assigned ad with a binary state variable $D \in \{0, 1\}$, where $D = 1$ if a user is exposed to the assigned ad, and $D = 0$ otherwise. While an eligible user assigned to ad Z will not necessarily be exposed to ad Z, the user will be exposed to ad Z_i if the user is exposed to the campaign at all.

2.1 Potential Outcomes: Extending the Rubin Causal Model

Following the Rubin (1974) causal model (RCM), we characterize the observed $Y_i^{(\text{obs})}$ as being one realization from a set of *potential outcomes*. Each potential outcome, $Y_Z^{(D)} = Y_i(D_i, Z_i)$, is a function of an ad treatment Z

[3] The individual creative elements could be defined as a set of ad attributes, but we are not studying those attributes explicitly.
[4] The audience is like a marketer's "target segment," but we avoid that phrase because the word "target" has a different meaning in the context of online ad delivery.
[5] Whether the platform randomly assigns ads to users at the start of the experiment or immediately before exposure (i.e., the real-time temporal ordering of this decision) does not matter, as long as the ad assignment probabilities are set before deciding which subset of users will be exposed to any ad in the campaign.
[6] An ad exposure means the platform presents the ad on the user's screen, regardless of whether the user actually laid eyes on the ad.

and exposure state D, but the exact nature of that function is unknown, heterogeneous, and non-stationary.[7] For instance, for $n_Z = 2$ and $Z \in \{z_1 = A, z_2 = B\}$, there are 4 possible states. One, $Y_A^{(1)}$ represents the potential outcome that would arise *if* the user were initially assigned to ad A *and if* the user were also exposed to ad A $(Z = A, D = 1)$. Another, $Y_B^{(0)}$, is the potential outcome *if* the user were assigned to ad $z_2 = B$, but were not exposed to that ad $(Z = B, D = 0)$. The user is endowed with all $2n_Z$ potential outcomes, but because the user will end up in exactly one of the $2n_Z$ possible (D, Z) states, the only potential outcome that is ever realized is $Y_i^{(\text{obs})}$. The others are hypothetical and counterfactual values. Without losing generality, we will use labels A and B to represent any two of the n_Z treatments and any one of the $\binom{n_Z}{2}$ possible pairwise comparisons.

We define an *effect* as a difference between *potential* outcomes. Still following the RCM, we are interested in three kinds of effects:

- $Y_Z^{(1)} - Y_Z^{(0)}$ is the difference between what the outcome *would have been if* the user were assigned and exposed to ad Z, and what the outcome *would have been if* that same user were assigned to the same ad Z but not exposed to it. We define ad Z's *lift*, λ_Z, to be the expected value of this effect across a subset of users. When defined over the entire audience, lift is an *Average Treatment Effect (ATE, Eq. 1)*. When lift is defined only among users who were actually exposed to the ad, it is an *Average Treatment Effect on the Treated (ATET, Eq. 2)*.

$$\lambda_Z^{\text{ATE}} = \mathbf{E}\left[Y_Z^{(1)} - Y_Z^{(0)}\right] \tag{1}$$

$$\lambda_Z^{\text{ATET}} = \mathbf{E}\left[Y_Z^{(1)} - Y_Z^{(0)} \,\middle|\, D = 1\right] \tag{2}$$

- $Y_A^{(1)} - Y_B^{(1)}$ is the difference between what the outcome *would have been if* the user were assigned and exposed to ad A, and what the outcome *would have been if* that same user were assigned and exposed to ad B. An analogous effect $Y_A^{(0)} - Y_B^{(0)}$ is the same difference in potential outcomes, but if users were *unexposed* to each ad.
- $\left(Y_A^{(1)} - Y_A^{(0)}\right) - \left(Y_B^{(1)} - Y_B^{(0)}\right)$ is the difference-in-differences in potential outcomes for any user. We define Δ_{AB}, the *A/B difference* between ads A and B, to be the expected value of this "diff-in-diff," which is the difference between the *lift* of ad A and the *lift* of ad B. The A/B difference can be defined over the entire audience (Δ_{AB}^{ATE}), or different subsets of users, such as for only exposed users ($\Delta_{AB}^{\text{ATET}}$).

$$\Delta_{AB}^{\text{ATE}} = \lambda_A^{\text{ATE}} - \lambda_B^{\text{ATE}} = \mathbf{E}\left[\left(Y_A^{(1)} - Y_A^{(0)}\right) - \left(Y_B^{(1)} - Y_B^{(0)}\right)\right] \tag{3}$$

$$\Delta_{AB}^{\text{ATET}} = \lambda_A^{\text{ATET}} - \lambda_B^{\text{ATET}} = \mathbf{E}\left[\left(Y_A^{(1)} - Y_A^{(0)}\right) - \left(Y_B^{(1)} - Y_B^{(0)}\right) \,\middle|\, D = 1\right] \tag{4}$$

The expected *unexposed* potential outcomes, $\mathbf{E}\left[Y_Z^{(0)}\right]$, play a central role in this paper because baseline propensities can vary across users with different (D, Z) states. In general, users may be in any given state for a non-random reason, controlled in large part by the platform. When ads are assigned to the audience randomly, then the

[7] We suppress the i subscript in $Y_Z^{(D)}$ to reduce notational clutter.

Table 1: Potential Outcomes under Proposed, RCT, and ITT Frameworks

	Proposed Framework				**RCT**		**ITT**	
Assignment:	$Z_i = A$	$Z_i = B$	$Z_i = C$	$\cdots$	$Z_i = \text{Ctrl}$	$Z_i = \text{Trt}$	$Z_i = \text{Ctrl}$	$Z_i = \text{Trt}$
Exposed ($D = 1$)	$Y_{\text{A}}^{(1)}$	$Y_{\text{B}}^{(1)}$	$Y_{\text{C}}^{(1)}$	$\cdots$	$Y_{\text{Ctrl}}^{(1)}$	$Y_{\text{Trt}}^{(1)}$	$Y_{\text{Ctrl}}^{(1)}$	$Y_{\text{Trt}}^{(1)}$
Not exposed ($D = 0$)	$Y_{\text{A}}^{(0)}$	$Y_{\text{B}}^{(0)}$	$Y_{\text{C}}^{(0)}$	$\cdots$	$Y_{\text{Ctrl}}^{(0)}$	$Y_{\text{Trt}}^{(0)}$	$Y_{\text{Ctrl}}^{(0)}$	$Y_{\text{Trt}}^{(0)}$

Note to Table 1: In the proposed framework, all potential outcomes are uniquely identified. In a RCT, $Y_{\text{Trt}}^{(0)} = Y_{\text{Ctrl}}^{(0)} = Y_{\text{Ctrl}}^{(1)}$. In an ITT, $Y_{\text{Ctrl}}^{(1)} = Y_{\text{Ctrl}}^{(0)}$. Equivalent values under those established frameworks are grouped.

expected unexposed potential outcomes do not depend on the user's assigned ad: $\mathbf{E}\left[Y_{\text{A}}^{(0)}\right] = \mathbf{E}\left[Y_{\text{B}}^{(0)}\right]$. A special case arises when *exposure* is randomly determined, as well. In this case, when probability of exposure $\mathbf{P}(D = 1)$ is the same for all users, the separate groups of exposed users and unexposed users are both representative random samples of the full audience. That is, the expected potential outcome when exposed to ad Zwould be the same for the users who were actually exposed, for those who were actually not exposed, and for the audience, $\mathbf{E}\left[\, Y_{\text{Z}}^{(1)} \mid D = 1 \,\right] = \mathbf{E}\left[\, Y_{\text{Z}}^{(1)} \mid D = 0 \,\right] = \mathbf{E}\left[Y_{\text{Z}}^{(1)}\right]$; similarly, for the other exposed potential outcome, $\mathbf{E}\left[\, Y_{\text{Z}}^{(0)} \mid D = 1 \,\right] = \mathbf{E}\left[\, Y_{\text{Z}}^{(0)} \mid D = 0 \,\right] = \mathbf{E}\left[Y_{\text{Z}}^{(0)}\right]$. Therefore, *in this case of randomized exposure*, $\lambda_{\text{Z}}^{\text{ATET}} = \lambda_{\text{Z}}^{\text{ATE}}$. Now, if we combine the random ad assignment and random exposure conditions, then $\mathbf{E}\left[\, Y_{\text{A}}^{(0)} \mid D = 1 \,\right] = \mathbf{E}\left[\, Y_{\text{B}}^{(0)} \mid D = 1 \,\right]$. So again, in this case of random exposure, the A/B difference reduces to a difference in only the *exposed* potential outcomes: $\Delta_{\text{AB}}^{\text{ATET}} = \mathbf{E}\left[Y_{\text{A}}^{(1)} - Y_{\text{B}}^{(1)} \,\middle|\, D = 1\right] = \mathbf{E}\left[Y_{\text{A}}^{(1)} - Y_{\text{B}}^{(1)}\right] = \Delta_{\text{AB}}^{\text{ATE}}$.

This extension to the standard potential outcomes framework, in which potential outcomes are defined for all combinations of assignment and exposure, nests extant randomized control trial (RCT) and intent-to-treat (ITT) designs, neither of which would permit the advertiser to estimate the difference in lifts between two ads (Eqs. 3 and 4) separately from lift for a single ad (Eqs. 1 and 2). We explain with the help of Table 1. Suppose we run an experiment with two "arms:" a treatment (focal) ad $Z = \text{Trt}$, and a control (placebo) ad $Z = \text{Ctrl}$.[8] Each user has $2n_Z = 4$ potential outcomes: $Y_{\text{Trt}}^{(1)}$, $Y_{\text{Ctrl}}^{(1)}$, $Y_{\text{Trt}}^{(0)}$, and $Y_{\text{Ctrl}}^{(0)}$. The potential effect of exposure to the focal ad is $\mathbf{E}\left[Y_{\text{Trt}}^{(1)} - Y_{\text{Trt}}^{(0)}\right]$. The basic idea of an RCT is to estimate this potential effect by estimating $\mathbf{E}\left[Y_{\text{Trt}}^{(1)} - Y_{\text{Ctrl}}^{(1)}\right]$ instead. The platform records $Y_i^{(\text{obs})} = Y_{\text{Trt}}^{(1)}$ for users exposed to $Z = \text{Trt}$, and because exposing a user to a placebo control is still exposing a user to *something*, it records $Y_i^{(\text{obs})} = Y_{\text{Ctrl}}^{(1)}$ for users exposed to $Z = \text{Ctrl}$. It is only through the assumptions of the RCT that exposure to the placebo control $Y_{\text{Ctrl}}^{(1)}$ is equivalent to the absence of exposure to the treatment $Y_{\text{Trt}}^{(0)}$. Also, if $Z = \text{Ctrl}$ is a true placebo, then $Y_{\text{Ctrl}}^{(1)}$ is assumed to equal $Y_{\text{Ctrl}}^{(0)}$, meaning that $Y_{\text{Trt}}^{(0)} = Y_{\text{Ctrl}}^{(0)} = Y_{\text{Ctrl}}^{(1)}$, and only two of the four potential outcomes are uniquely defined (Table 1, RCT grouping). If we accept the same RCT assumptions that allow us to estimate a causal effect by comparing responses to a treatment and control ad, we cannot also separately identify a within-ad and between-ad effect in the same framework.

An ITT design, as in Gordon et al. (2019), does not solve this problem. As in the RCT, users are randomly

[8] Because we use 0/1 to indicate exposure states, we opt for labels for ads (e.g., A/B, Trt/Ctrl).

assigned to $Z = \text{Trt}$ or $Z = \text{Ctrl}$ arms, and all $Z = \text{Trt}$ users are *intended* to be treated. But in an ITT design, some treatment users are exposed ($Z = \text{Trt},\ D = 1$; *compliant*) to the treatment, while the others are unexposed ($Z = \text{Trt},\ D = 0$; *noncompliant*).[9] Depending on D, the platform records $Y_i^{(\text{obs})}$ to be either $Y_{\text{Trt}}^{(1)}$ or $Y_{\text{Trt}}^{(0)}$ from users in the treatment arm. Exposure among users assigned to the control arm is not considered explicitly, so they could either be exposed to a placebo control ($Y_i^{(\text{obs})} = Y_{\text{Ctrl}}^{(1)}$), or perhaps no ad at all ($Y_i^{(\text{obs})} = Y_{\text{Ctrl}}^{(0)}$). But exposure among the treated users is not necessarily random, and permits those "intended-to-treat-but-unexposed" users' potential outcomes ($Y_{\text{Trt}}^{(0)}$) to be different from $Y_{\text{Ctrl}}^{(0)} = Y_{\text{Ctrl}}^{(1)}$. Therefore, ITT uses three of the four potential outcomes: $Y_{\text{Trt}}^{(1)}$ for the compliant, treatment users; $Y_{\text{Trt}}^{(0)}$ for the noncompliant treatment users, and $Y_{\text{Ctrl}}^{(1)} = Y_{\text{Ctrl}}^{(0)}$ for the control/placebo users (Table 1, ITT grouping). Importantly, unless compliance were random, $\mathbf{E}\left[Y_{\text{Trt}}^{(1)} - Y_{\text{Trt}}^{(0)}\right]$ would not be equal to $\mathbf{E}\left[Y_{\text{Trt}}^{(1)} - Y_{\text{Ctrl}}^{(1)}\right]$, so comparison of treated and exposed users to a placebo group is not an acceptable alternative for describing the lift of an ad treatment.

In this paper, we need to think about assignment as a completely different concept from exposure. Because we want to separate the effect of exposure to an ad (e.g., $\lambda_{\text{A}} = \mathbf{E}\left[Y_{\text{A}}^{(1)} - Y_{\text{A}}^{(0)}\right]$, a within-ad difference) from the effect of the initial ad assignment to an ad (a between-ad difference, or interaction like $\Delta_{\text{AB}} = \lambda_{\text{A}} - \lambda_{\text{B}}$) , we need access to all $2n_Z$ potential outcomes, which is something the RCM does not explicitly provide. That way we can compare any possible combinations of ad-assignment-exposure states by just taking a difference between these potentials. Values of λ_{Z} and Δ_{AB} are simple differences and differences-in-differences between the rows and columns in Table 1. Further, each comparison can be considered over any subgroup of users (e.g., exposed users, targeted users, whole audience). These differences in potentials are nevertheless just theoretical constructs to be estimated from observed data, and they may require assumptions about the randomization between ad assignment and exposure states in the experimental design.

2.2 Targeting, Effect Estimation, and Experimental Design

Even with access to all potential outcomes for all ads and exposure states, the realities of targeting policies and experimental design dictate which causal effects can be estimated from the data. Within this generalized potential outcomes framework, we can formally define the inference problem at hand. We begin with the relationships among the user types, assigned ads, and targeting algorithm. Let $X_i \in \{x_1, x_2, \dots\}$ represent user i's *user type*, which includes both *observable* traits that define the advertiser's pre-specified audience and dimensions across which experimental results will be summarized; and *unobserved* traits (all other information about user i that is stored on the platform). Still, regardless of their type, each user is randomly assigned to ad treatment group Z_i.

Fig. 1 presents abstractions of two possible experimental designs for estimating causal effects of ads from a campaign with three ads, $Z \in \{A, B, C\}$. For both approaches, the audience is randomly partitioned into three

[9] To maintain consistency with standard terminology, a user who is *exposed* to the assigned treatment is *compliant* with that treatment, regardless of the reason for exposure. A user *assigned* to the treatment is *intended* to be treated, regardless of the true intentions of the experimenter.

"audience-ad squares," each corresponding to an ad. Users are represented by circles, and user types X_i are distinguished by color and vertical position of those circles. Because ads are randomly assigned, each audience-ad square contains the same mix of user types (mix of colors). Only after the randomization of an audience into separate groups of users assigned to each ad does the *targeting algorithm* go to work. We characterize the targeting algorithm as a function, $\tau: (Z_i, X_i, \Omega) \mapsto \{0, 1\}$, where user i is either targeted ($\tau = 1$), or untargeted ($\tau = 0$). The inputs to this targeting algorithm function are Z_i, X_i, and a generic placeholder Ω, containing any other information the algorithm has at its disposal from across the platform.[10] Because the internal operations of the targeting algorithm are complex, proprietary, and unobservable (as if inside a black box), we treat the targeting function *as if* it were a conditionally random process from the point of view of the advertiser, with *targeting probabilities* $\mathbf{P}(\tau = 1 \mid X, Z)$.

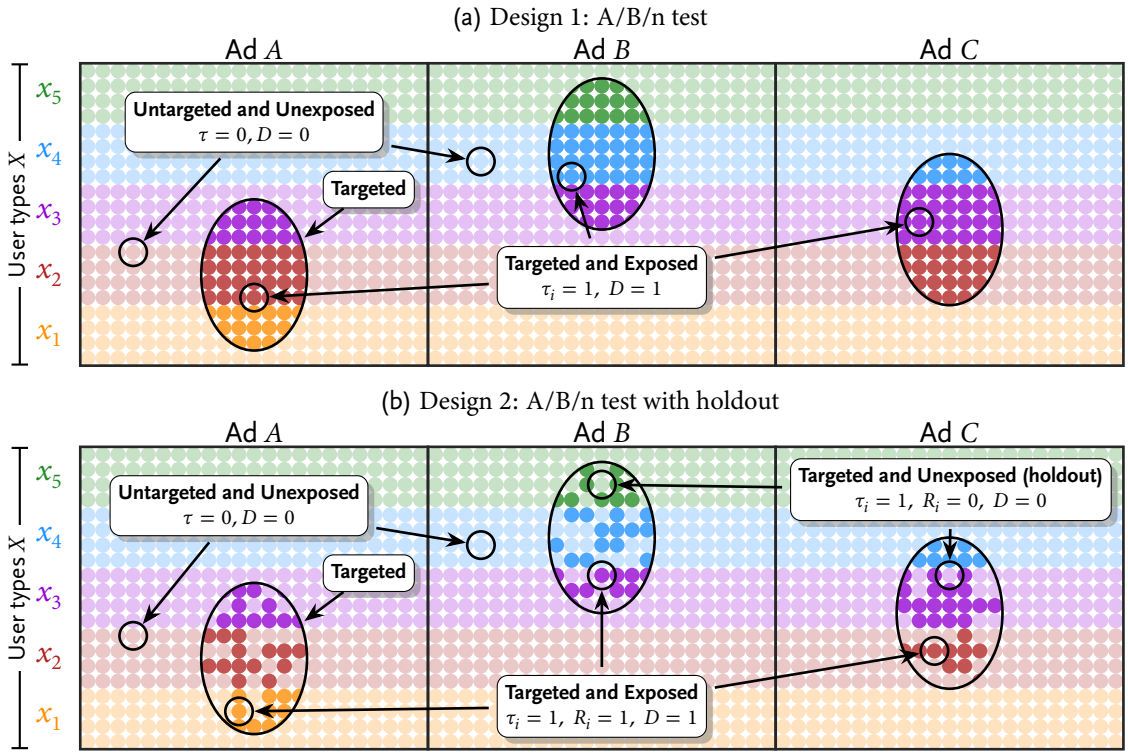
In Fig. 1, the targeted ($\tau = 1$) users assigned to ad Z sit inside the "targeting ovals" within each audience-ad square. The remaining users outside the targeting ovals are untargeted ($\tau = 0$). The mix of *targeted* users' types (colors inside the oval) is different from the mix of types in the entire ad-audience square. Because the algorithm considers a user's assigned ad when deciding if that user should be targeted (i.e., the targeting probability depends on Z), the resulting mix of user types among targeted users varies across ads. Thus, Fig. 1 visualizes *divergent delivery*: the mixes of targeted user types (colors within and vertical positions of each targeting oval) are different for each ad. Being targeted, however, is a necessary but not sufficient condition for being exposed. Targeted users may be exposed to the assigned ad ($D = 1$, bright colored circles), or be unexposed ($D = 0$, dimmed colored circles). Whether a targeted user is exposed to the assigned ad depends on the design of the experiment, two of which we discuss now.

Design 1: the A/B/n test The first design with these features is the **A/B/n test** (Fig. 1a). Intended for comparisons across distinct ad creatives, the A/B/n test design lets the advertiser compare outcomes from users targeted and exposed to ad A, to outcomes from users targeted and exposed among ads B or C. Targeted users are exposed to their assigned ads (meaning that $\tau = D$), and the platform reports outcomes that are aggregated over users targeted with a given ad. Ad C could be (but does not have to be) a control ad, like a baseline ad created by the advertiser as a reference point, or a placebo ad that is unrelated to the campaign (e.g., a public service announcement). However, under our framework, *and reflecting how targeting algorithms operate in practice*, a control (or placebo) ad is just another ad that the platform targets to different users, regardless of the ad's role in the advertiser's experimental design.

This A/B/n test setup presents two immediate and practical concerns regarding interpretation of comparisons across ads.

[10] Ω includes all other users' types and histories, as well as items like when the ad would be presented (e.g., seasonality, day, time), the state of the auction (e.g., competition, bids), and parameters of the campaign.

Figure 1: A/B Test and A/B Test with Holdout Designs for Online Ad Experiments



Note to Figure 1: Each circle is a user in the audience, and colors and vertical positions represent unobserved user types. Users are randomly assigned to ad *A*, *B* or *C*, so the mix of user types for the audience is the same in all three ad-audience squares. Users inside each targeting oval are targeted with their assigned ad, and the targeted mix in each oval differs from the audience mix. Since the algorithm can deliver each ad to different users, the mix of user types in among the targeted users also varies across ads (color and vertical positions differ across targeting ovals). Bright dots are users who are exposed to the corresponding ad, while dim dots are unexposed. In a A/B/n test (Fig. 1a), all targeted users are exposed, while in a A/B/n test with holdout (Fig. 1b), users are randomly assigned to a treatment arm or holdout arm.

Concern 1: Targeted v untargeted mixes. In an A/B/n test design, targeting is not random. The distributions of potential outcomes *if they were to have been* untargeted will be different, so the targeted and untargeted groups are not comparable. That is, the users who ended up being targeted (inside the ovals in Fig. 1a) and the users who ended up being untargeted (outside the ovals) may have different expected potential outcomes for the unexposed state: $\mathbf{E}\left[Y_Z^{(0)} \mid \tau = 1 \right] \neq \mathbf{E}\left[Y_Z^{(0)} \mid \tau = 0 \right]$. The non-random targeting algorithm creates a confound that interferes with the advertiser's goal of inferring the *incremental* impact of exposure to the ad itself, separately the effect from how the algorithm targets the users to see the ad. If the observed data used to estimate $\mathbf{E}\left[Y_Z^{(1)} \middle| \tau = 1\right]$ and $\mathbf{E}\left[Y_Z^{(0)} \middle| \tau = 0\right]$ are collected from sets of targeted and untargeted users with different mixtures of types, the advertiser cannot know if an estimate of $\mathbf{E}\left[Y_Z^{(1)} \middle| \tau = 1\right] - \mathbf{E}\left[Y_Z^{(0)} \middle| \tau = 0\right]$ is measuring the incremental value of the ad creative, the impact of how the algorithm decides which users are targeted, or a combination of the two. Non-random targeting means that the results of the experiment will not reflect the true lift of the ad for the audience.

Concern 2: A vs B mixes. The second concern with the A/B/n test design is that the targeted groups of users are not comparable across ads. The algorithm targets ad A differently from how it targets ad B, so ad A's distribution of potential outcomes among targeted users is not the same as B's (vertical positions of the targeting ovals in Fig. 1a). While the estimates of the quantities $\mathbf{E}\left[Y_{\text{A}}^{(1)} - Y_{\text{A}}^{(0)} \,\middle|\, \tau = 1\right]$ and $\mathbf{E}\left[Y_{\text{B}}^{(1)} - Y_{\text{B}}^{(0)} \,\middle|\, \tau = 1\right]$ can *separately* be interpreted as causal effects, the difference in these lifts, $\mathbf{E}\left[Y_{\text{A}}^{(1)} - Y_{\text{A}}^{(0)} \,\middle|\, \tau = 1\right] - \mathbf{E}\left[Y_{\text{B}}^{(1)} - Y_{\text{B}}^{(0)} \,\middle|\, \tau = 1\right]$, cannot reflect the causal effect of assignment of ad A vs B. Advertisers cannot distinguish between the true difference in effect between the creative elements of ads A and B, and the effect from the targeting algorithm's selection of users to see each ad. As long as targeting creates one mix of users to see ad A and a different mix to see ad B, causal inference about the A/B difference in lifts, even among targeted users, is in jeopardy.[11]

Concern 2 is at the heart of why the common practice of comparing results of a focal ad of a campaign with a placebo control fails to reveal a true causal effect of exposure to that ad. As discussed in Sec. 2.1, when making this comparison in a standard RCT, the advertiser is assuming that observations from users exposed to the placebo ad C can substitute for unobserved users assigned to, but not exposed to, ad A ($Y_{\text{C}}^{(1)} = Y_{\text{A}}^{(0)}$), to make $\mathbf{E}\left[Y_{\text{A}}^{(1)} - Y_{\text{C}}^{(1)} \,\middle|\, \tau = 1\right] = \mathbf{E}\left[Y_{\text{A}}^{(1)} - Y_{\text{A}}^{(0)} \,\middle|\, \tau = 1\right]$. Concern 2 explains why we cannot maintain that assumption under non-random exposure: the mixes of types of targeted users are different for ad A and the placebo control C.

Design 2: Fig. 1b illustrates an **A/B/n test with holdout design**, (sometimes known as a split lift test). Conditional on being targeted ($\tau = 1$), users are randomized into one of two "arms " of the design: (1) a *treatment arm* ($R = 1$) whose users are exposed ($D = 1$, bright circles inside the targeting oval); and (2) a *holdout arm* ($R = 0$) whose users are unexposed ($D = 0$, dimmed circles inside the targeting oval). For each ad, $Y_{\text{Z}}^{(1)}$ is observed for users in the treatment arm, and $Y_{\text{Z}}^{(0)}$ is observed for users in the holdout arm.

Concern 1 is partially resolved by Design 2. Design 2's additional randomization step among targeted users is a recent innovation in online experimentation (Johnson et al. 2017a). Platforms that implement this design are essentially running "two-armed mini-RCTs" among only the targeted users (Gordon et al. 2019). Because targeted users are randomly assigned to treatment and holdout arms (the bright circles are randomly selected among the all of the circles inside the ovals in Fig. 1b), this estimate of $\mathbf{E}\left[Y_{\text{Z}}^{(1)} - Y_{\text{Z}}^{(0)} \,\middle|\, \tau = 1\right]$ does have a causal interpretation for the set of targeted users. This design is akin to an ITT design where the targeting process is flagging users who are "intended to be treated." But when targeting is non-random (mixes of colors in the targeting oval are different from the mixes in the entire audience square), then $\mathbf{E}\left[Y_{\text{Z}}^{(1)} - Y_{\text{Z}}^{(0)} \,\middle|\, \tau = 1\right] \neq \mathbf{E}\left[Y_{\text{Z}}^{(1)} - Y_{\text{Z}}^{(0)}\right]$. Therefore, we cannot say that Concern 1 is resolved completely.

Recognizing the importance of measuring incremental effects, some publishers have already deployed tools that resolve Concern 1. For example, the Johnson et al. (2017a) "ghost ads" method described in Sec. 1 has been

[11] Even the notion of the difference between lifts of ads targeted to different groups has a strained interpretation. Rather than there being a single targeted group of users, there are two different groups, A and B, that differ in unobserved ways.

implemented in practically equivalent forms by Google and Facebook. These tools allow platforms to create the needed experimental variation and to report the appropriate comparison of average outcomes that best reflects true counter-factual comparison of potential outcomes that defines an single ad's lift.

But Concern 2 remains unresolved. Randomly deciding whether a targeted user will be exposed (bright circles inside the ovals in Fig. 1b) or unexposed (dimmed circles inside the ovals) does not resolve Concern 2 because the mix of types among targeted users is different for each ad (the vertical positions of the ovals in Fig. 1b). Even though each lift was estimated from a two-armed RCT experiment, the effects are still computed from different mixes of targeted users, ($R = 1 \mid \tau = 1$ vs. $R = 0 \mid \tau = 1$). As a result, the difference in lift across ads, $\mathbf{E}\left[Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{A}}^{(0)} \,\middle|\, \tau = 1\right] - \mathbf{E}\left[Y_{\mathrm{B}}^{(1)} - Y_{\mathrm{B}}^{(0)} \,\middle|\, \tau = 1\right]$ does not have a causal interpretation because of the confound between the targeted mix selection and the ad. Even worse, *the advertiser will not see this confound and cannot detect how large of a problem this confound will be.*[12]

The advertiser has no immediate and existing remedy for this second concern. For the rest of this paper, we will consider only the A/B/n test with holdout (Fig. 1b), because at least it resolves part of Concern 1. Later, we will present a remedy that the *platform* could implement to address Concern 2.

2.3 Remarks on Mechanisms of Ad Exposure and Delivery

Availability bias and compliance. In addition to targeting (τ) and random holdout (R), there is a third determinant of exposure: *user availability* (V), also known as activity bias (Lewis et al. 2011). The experiment runs for a predetermined time, and some users who might otherwise be part of the subject pool will be excluded simply because they were not "available" (e.g., did not log into or spend enough time on platform). Let $V = 1$ denote users who are available, and thus are eligible to be targeted, while $V = 0$ indicates users are not eligible to be targeted. In order to be exposed, a user must be available, targeted, and, in the case of an A/B/n test with holdout, assigned to the treatment arm. Consequently, the probability of exposure, conditional on random assignment to Z_i, is

$$\mathbf{P}(\, D = 1 \mid Z_i, X_i, \Omega \,) = \mathbf{P}(\, R = 1 \mid \tau = 1 \,)\mathbf{P}(\, \tau = 1 \mid Z_i, X_i, \Omega, V = 1 \,)\mathbf{P}(\, V = 1 \mid X_i \,) \tag{5}$$

Using the language of experimental design and causal inference, a user who is assigned to Z_i is "in compliance" if the user is also exposed to Z_i. Under the RCM and its extensions, compliance is usually thought of as one-dimensional. But in our framework for online ad experiments, the three factors in Eq. 5 are three separate types of compliance: (1) random compliance through R; (2) *algorithmic compliance* from the targeting process τ; and (3) *user compliance* based on user availability V.

[12] Some platforms provide experimental results that are broken down by coarse demographic groups (e.g., gender, age). In our framework, those kinds of demographic subgroups are *observable* and *define different audiences.* These are not the *unobserved* elements of X_i that determine which users to *target within an audience.* In this paper, we are only referring to the confound generated by *unobserved* traits.

In Eq. 5, $\mathbf{P}(\, R = 1 \mid \tau = 1\,)$ is *ignorable* with respect to Z_i and X_i; the probability a targeted user is randomly assigned to the treatment arm is the same for all ads and user types, by construction of A/B/n test with holdout. But algorithmic and user compliances are *non-ignorable* when the targeting and user availability probabilities depend on user type X_i and/or ad Z_i. Algorithmic compliance is non-ignorable when targeting is non-random: $\mathbf{P}(\, \tau = 1 \mid V = 1, X_i, Z_i, \Omega\,)$. But we will assume that availability compliance *is ignorable*, where $\mathbf{P}(\, V = 1 \mid X_i\,) = \mathbf{P}(V = 1)$ for all users. Nevertheless, this framework can support future research on the drivers of "availability bias" and selection effects by relaxing this ignorability assumption on V. But for our purposes, we will let exposure D depend on Z_i and X_i only through τ.

What do targeted but unexposed users see? Only the $V = 1$ and $\tau = 1$ users are part of the experiment, so only their data (aggregated) is included in the reports to the advertiser. While the $R = 1$ users are exposed to their assigned ad Z, the $R = 0$ users are not. Instead, the platform delivers those "holdout" users a "shadow control" ad $S_{Z,i}$ that is determined by a function $S(Z_i, X_i, \Omega)$ that returns the second place ad that would have been shown to user i in that exact time and context, if Z_i did not exist.

Diagramming the path to exposure. The tree in Fig. 2 ties the exposure, targeting, and compliance processes together. It provides an annotated tour of how exposure to one of the experimental treatment ads requires the user to pass through three "filters": availability, targeting, and random assignment to the treatment arm.

3 Defining causal effects with targeting and heterogeneity

Now that we have explained the intuition behind how targeting algorithms that exploit users' heterogeneity in responsiveness to ads can yield problematic comparisons of effects between ads, we can formalize that issue mathematically. By allowing for the probability of exposure (Eq. 5) to depend on both the ad creative and user type, we extend the existing approach in the literature. This section extends our notation and definitions; the analytic and simulation results will follow in Sec. 4.

3.1 Lift and targeting with heterogeneous users

A user type X_i (introduced in Sec. 2.2) encompasses all of the user characteristics relevant for describing user preferences and propensities (i.e., users' behavioral propensities, whether exposed or unexposed), as well as for the platform's decisions about targeting and experimental design. We define $\gamma_X = \mathbf{P}(X)$ to be the proportion of type X users in the audience, or equivalently, the prior probability that a randomly selected member of the audience has type X. Further, let $\mathbf{E}\left[\, Y_Z^{(1)} \mid X \,\right]$ and $\mathbf{E}\left[\, Y_Z^{(0)} \mid X \,\right]$ be the expected potential outcomes among users with type X. The lift of ad Z for users of type X, λ_{XZ}, is the expected difference in these potential outcomes, or type-specific lift of an ad,

$$\lambda_{XZ} = \mathbf{E}\left[\, Y_Z^{(1)} - Y_Z^{(0)} \,\middle|\, X \,\right] \tag{6}$$

Figure 2: Conceptual Tree Framework for A/B/3 Test with Algorithmic Ad Targeting

The advertiser defines the **audience** using criteria available on the platform.

All users in the audience are randomly **assigned** to a treatment Z_i.

Unavailable users ($V=0$) cannot see any ads. **Available** users ($V=1$) are eligible to be targeted.

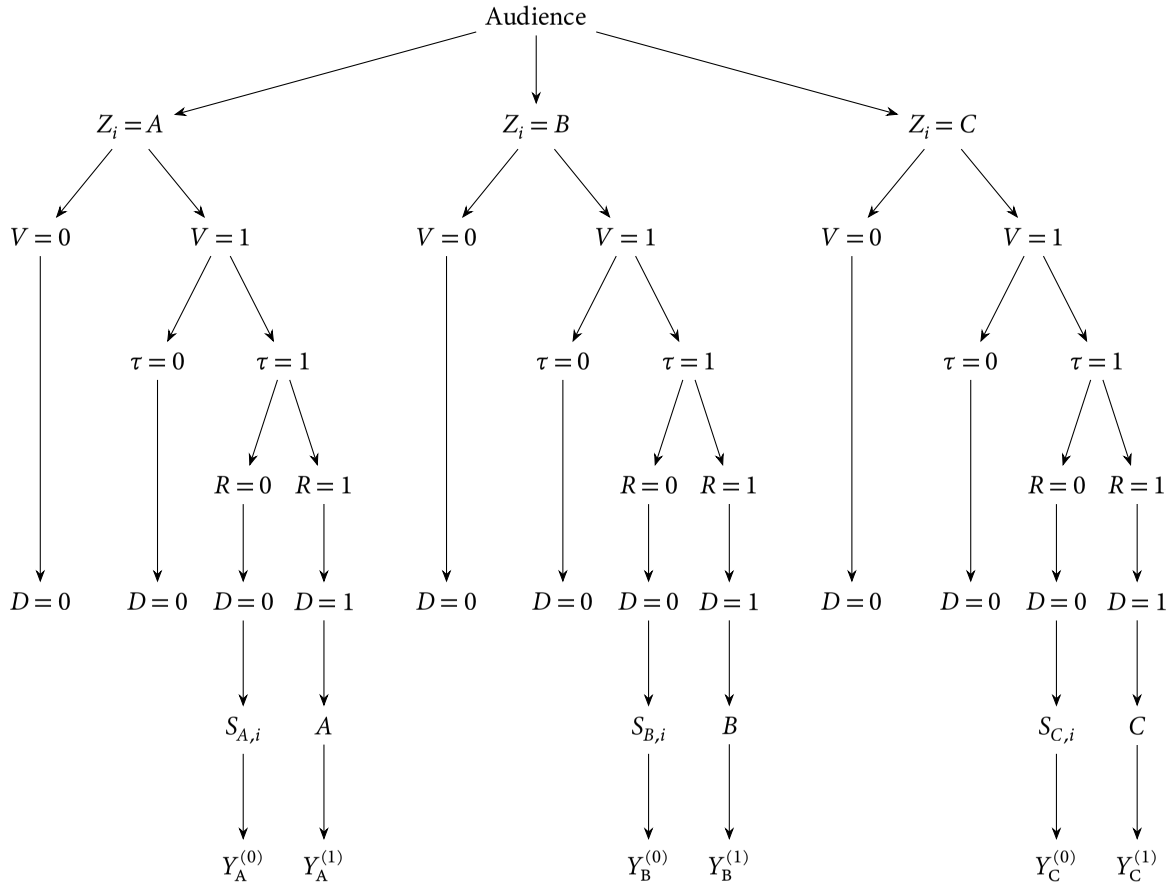
Available **untargeted** users ($V=1, \tau=0$) will see ads that are unassociated with the focal campaign.

Available **targeted** users ($V=1, \tau=1$) are randomized into either the treatment arm ($R=1$) or the holdout arm ($R=0$).

A user is **exposed** to its assigned ad if and only if $D=\tau \cdot V \cdot R=1$.

Targeted exposed users ($\tau=1, D=1$) **see** the assigned ad. Targeted unexposed users ($\tau=0, D=0$) see the shadow control ad $S_{Z,i}$.

The **recorded outcome** for a user exposed to Z_i is $Y_Z^{(\text{obs})} = Y_Z^{(1)}$. The recorded outcome for an unexposed user who sees the alternative shadow control ad $S_{Z,i}$ is $Y_Z^{(\text{obs})} = Y_Z^{(0)}$.



Note to Figure 2: This tree illustrates how the platform in our framework subsets users, each randomly assigned ad $Z \in \{A, B, C\}$, into those who are targeted and exposed, targeted but unexposed, and untargeted. Each user is endowed with all 6 potential outcomes. $Y_Z^{(1)}$ is recorded for targeted, treatment arm users who see their assigned ad, while $Y_Z^{(0)}$ is recorded for targeted, holdout arm users who see the shadow control ad. The shadow control will be different for each user. No experimental data is recorded for unavailable or untargeted users. In practice, the temporal ordering of the levels of the tree may be different.

The expected *aggregate* lift $\lambda_{\mathrm{Z}}^{\mathrm{ATE}}$ (marginal across ads) and the expected aggregate A/B difference $\Delta_{\mathrm{AB}}^{\mathrm{ATE}}$ from Eqs. 1 and 3 can be expressed as mixtures of their type-specific counterparts, with γ_{X} as the mixture weights:

$$\lambda_{\mathrm{Z}}^{\mathrm{ATE}} = \sum_{\forall X} \lambda_{\mathrm{XZ}} \gamma_{\mathrm{X}} \tag{7}$$

$$\Delta_{\mathrm{AB}}^{\mathrm{ATE}} = \sum_{\forall X} \gamma_{\mathrm{X}} \left(\lambda_{\mathrm{XA}} - \lambda_{\mathrm{XB}} \right) \tag{8}$$

Because we conceptualize the targeting algorithm as probabilistic, we define the *targeting probability* for a randomly chosen user with type X and who is assigned to ad Z, to be

$$\Phi_{\mathrm{XZ}} = \mathbf{P}(\tau = 1 \mid X, Z) \tag{9}$$

Then, the marginal targeting probability for all users who were assigned to ad Z is a mixture of the type-specific targeting probabilities summed over the prior distribution of user types.

$$\Phi_{\mathrm{Z}} = \mathbf{P}(\tau = 1 \mid Z) = \sum_{\forall X} \Phi_{\mathrm{XZ}} \gamma_{\mathrm{X}} \tag{10}$$

The campaign-level aggregate probability that *any* user in the audience is targeted, is a mixture of the marginal ad-specific targeting probabilities, weighted by the initial random assignment probabilities, ζ_{Z}.

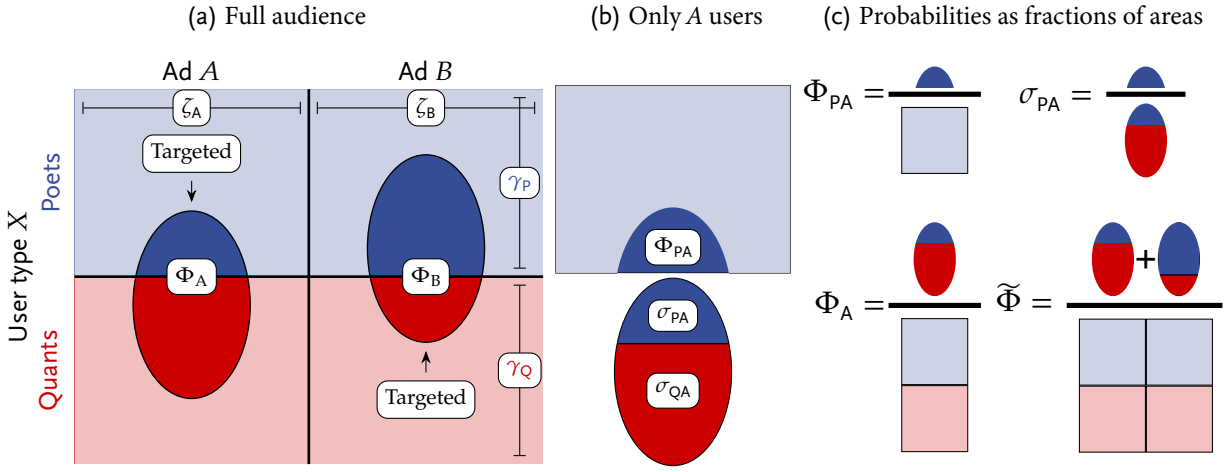
$$\widetilde{\Phi} = \mathbf{P}(\tau = 1) = \sum_{\forall Z} \Phi_{\mathrm{Z}} \zeta_{\mathrm{Z}}, \tag{11}$$

The distribution of user types among only targeted users who were assigned to ad Z is different from distribution of user types among the audience. While γ_{X} represents the prior mixture of user types among all users in the audience, we define σ_{XZ} to be the posterior mixture only among targeted users. Applying Bayes' Theorem,

$$\sigma_{\mathrm{XZ}} = \mathbf{P}(X \mid \tau = 1, Z) = \frac{\mathbf{P}(\tau = 1 \mid X, Z)\mathbf{P}(X)}{\mathbf{P}(\tau = 1 \mid Z)} = \frac{\Phi_{\mathrm{XZ}}}{\Phi_{\mathrm{Z}}} \gamma_{\mathrm{X}} \tag{12}$$

Fig. 3 illustrates these definitions of Φ_{XZ}, Φ_{X}, ζ_{Z}, γ_{X} and σ_{XZ}, using an example with two ads, $z_1 = A$ and $z_2 = B$, and two user types, $x_1 = P$ and $x_2 = Q$. For expositional clarity we will name these types Poets and Quants. In Fig. 3a, areas of the two dark "targeting ovals" in each column are the same proportions as the areas of their respective columns ($\Phi_{\mathrm{A}} = \Phi_{\mathrm{B}}$). But in this example, Poets who are randomly assigned to ad A ("A-Poets ") are less likely to be targeted than a randomly chosen user assigned to A ("A-user"). That is, $\Phi_{\mathrm{PA}} < \Phi_{\mathrm{A}}$. Visually, the proportion of the left blue square in Fig. 3a that is inside the targeting oval (Φ_{PA}, also in the top of Fig. 3b) is smaller than the proportion of the left *column* of Fig. 3a that is inside (Φ_{A}). From Eq. 12, $\sigma_{\mathrm{PA}} < \gamma_{\mathrm{P}}$, so the blue proportion of the A oval in Fig. 3b (bottom) is smaller than the blue proportion of the entire audience (Fig. 3a, full grid). On the other hand, B-Poets are *more* likely to be targeted than B users overall ($\Phi_{\mathrm{PB}} > \Phi_{\mathrm{B}}$), so $\sigma_{\mathrm{PB}} > \gamma_{\mathrm{P}}$. Therefore, $\sigma_{\mathrm{PA}} < \gamma_{\mathrm{P}} < \sigma_{\mathrm{PB}}$. We distinguish between these two effects: (1) targeting by user types overall occurs

Figure 3: Definitions of Φ_{XZ}, Φ_Z, $\widetilde{\Phi}$, and σ_{XZ} for a Two-Ad Experiment and a Two-Type Audience.



Note to Figure 3: Areas are proportional to numbers of users, so ratios of areas represent probabilities. Fig. 3a represents the audience, with Poets in blue on top and Quants in red on bottom. Row heights are proportional to mixture proportions γ_P and γ_Q. Each column is a randomly assigned ad, with widths proportional to assignment probabilities ζ_A and ζ_B. Targeted users are contained in the darkened "targeting ovals." Marginal targeting probabilities Φ_A and Φ_B are proportions of columns that are within their respective ovals. Fig. 3b, top: Φ_{PA} is the probability that a A-Poet is targeted (proportion of the A-Poet quadrant inside the oval). Fig. 3b, bottom: Posterior probabilities σ_{PA} and σ_{QA} are proportions of users *targeted* with A who are Poets and Quants, respectively. Fig. 3c defines probabilities as fractions of not-to-scale areas.

as the posterior mixture probability deviates from the prior mixture ($\sigma_{XZ} \neq \gamma_X$); and (2) *divergent delivery* occurs when the mix of users targeted with one ad does not resemble the mix targeted with another ad ($\sigma_{XA} \neq \sigma_{XB}$).

3.2 Characterizing a campaign's targeting policy

Building off of the targeting probabilities illustrated in Fig. 3, we now characterize these values with a parsimonious set of ratios. The following ratios define relationships among targeting probabilities between ads (α_τ), between user types (π_τ), and their interactions (ρ_τ).[13].

$$\alpha_\tau = \frac{\Phi_A}{\Phi_B} = \frac{\gamma_P \Phi_{PA} + \gamma_Q \Phi_{QA} + \cdots}{\gamma_P \Phi_{PB} + \gamma_Q \Phi_{QB} + \cdots} \qquad \text{(marginal ad targeting)} \qquad (13)$$

$$\pi_\tau = \frac{\Phi_P}{\Phi_Q} = \frac{\zeta_A \Phi_{PA} + \zeta_B \Phi_{PB} + \cdots}{\zeta_A \Phi_{QA} + \zeta_B \Phi_{QB} + \cdots} \qquad \text{(marginal user targeting)} \qquad (14)$$

$$\rho_\tau = \frac{\Phi_{PA}}{\Phi_{PB}} \Big/ \frac{\Phi_{QA}}{\Phi_{QB}} = \frac{\sigma_{PA}}{1-\sigma_{PA}} \Big/ \frac{\sigma_{PB}}{1-\sigma_{PB}} \qquad \text{(divergent delivery)} \qquad (15)$$

These ratios answer specific questions about a platform's algorithmic targeting policies in terms of pairwise comparisons across user types and ads.[14] How much more does the algorithm target one type of users over another, on average? If $\pi_\tau > 1$, the targeting algorithm is more likely to target Poets than Quants with the

[13]The second equality in Eq. 15 comes from substitution of each σ_{XZ} after solving Eq. 12

[14]While we can generalize these ratios to any number of ads and user types using a matrix of pairwise relationships, we will keep things simple by relying on a 2 ads × 2 user types design.

campaign overall. How much more does the algorithm target users with one ad over another? If $\alpha_\tau > 1$, the algorithm targets users assigned to ad A more than ad B. How differently does the algorithm target ads A and B to different user types? This interaction is *divergent delivery*, which we operationalize by the odds ratio ρ_τ. If $\rho_\tau > 1$, then the targeting favors those Poets assigned to A and those Quants assigned to B *even more than* whatever the marginal ratios α_τ and π_τ alone would have indicated.

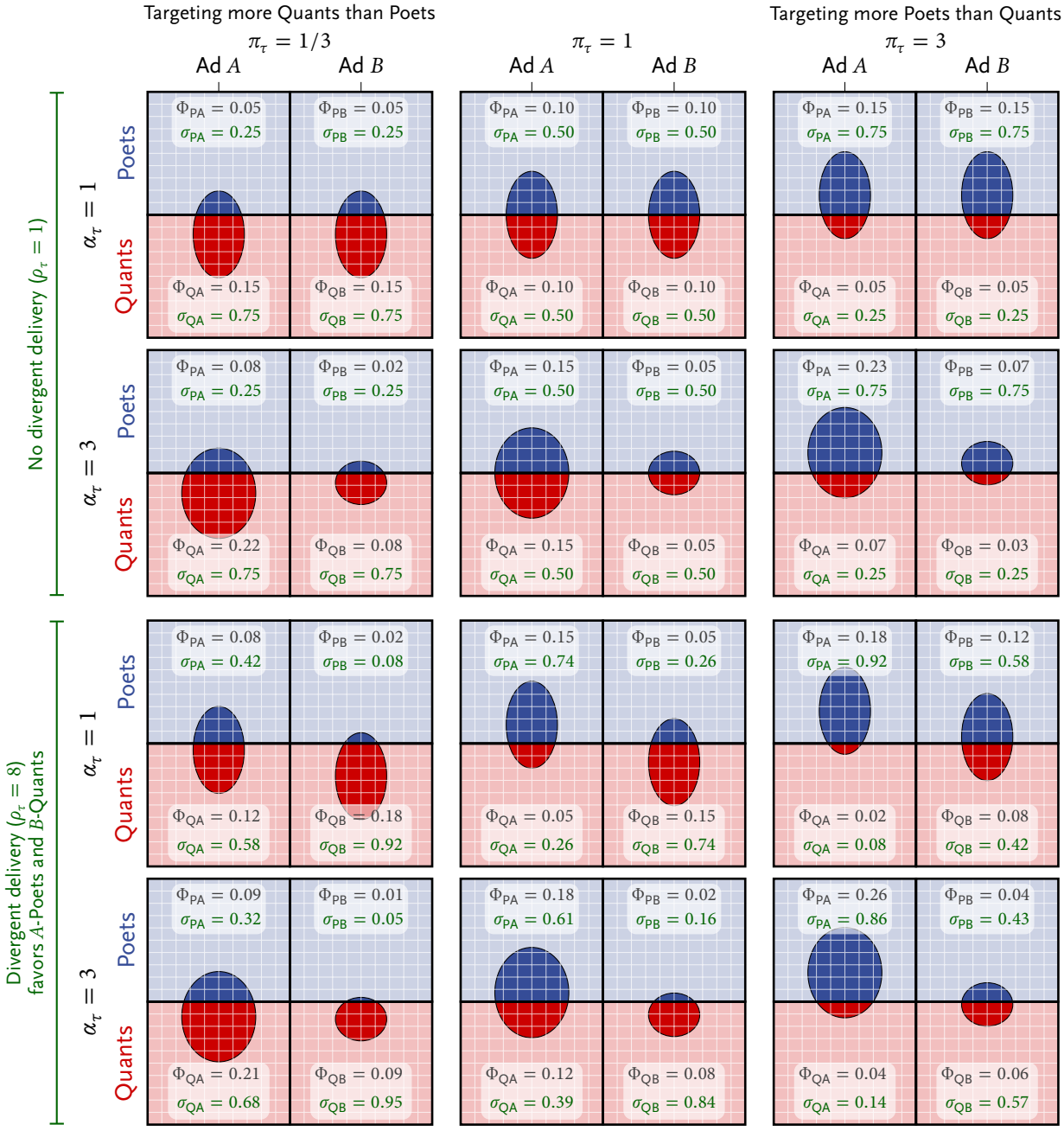
Figure 4 illustrates how different combinations of these three ratios' values correspond to distinct targeting policies; the Appendix contains a more mathematical treatment. Each set of ratios π_τ, α_τ, and ρ_τ defines a 2×2 panel in Fig. 4, with quadrants similar to Fig. 3a. Panels vary only by targeting policies for each X, Z pair. Visually inspecting how the colored portion is distributed across the areas of the ovals tells the story of how targeting efforts are proportionally distributed across users types and ads. If $\alpha_\tau > 1$, the area of the A oval is larger than the area of the B oval (Fig. 4, rows 2 and 4). If $\pi_\tau > 1$, the total blue area inside the A and B ovals increases as the ovals shift up *together* (right column). If $\rho_\tau > 1$, the blue area of the A oval and the red area of the B oval both increase as the vertical positions of the ovals *separate* (rows 3 and 4).

The relationship between divergent delivery (ρ_τ) and the posterior mixtures of targeted users (σ) in Eq. 15 is a key insight of this paper. We explain with two examples from Fig. 4. The top right panel describes a targeting policy that is equally likely to target ad A and ad B users ($\alpha_\tau = 1$; the targeting ovals have equal area), is three times more likely to target Poets than Quants ($\pi_\tau = 3$; more blue area than red area across both ovals), and creates a mix of targeted users that is the same across ads A and B (no divergent delivery, $\rho_\tau = 1$; the two ovals have the same color mix and vertical position). Even when there is no divergent delivery ($\rho_\tau = 1$ and $\sigma_{\text{PA}} = \sigma_{\text{PB}}$), the mix among all targeted users will not necessarily be the same as the mix in the audience ($\sigma_{\text{PA}} = \sigma_{\text{PB}} \neq \gamma_{\text{P}}$. The targeted mix will be the same as the audience mix only when $\rho_\tau = 1$ *and* $\pi_\tau = 1$. Also, note that when an algorithm *without* divergent delivery targets ad A more than ad B in aggregate (as in row 2 with $\alpha_{\text{Y}} = 3$), the probability an A-Poet is targeted (Φ_{PA}; the total dark blue area in the A quadrants) will be higher than when it targets A and B equally (row 1; $\alpha_\tau = 1$), but the mix of Poets among the targeted A-users (σ_{PA}; the proportion of the A targeting oval that is blue) remains the same for both values of α_{Y}. In the bottom center panel of Fig. 4, the algorithm is more likely to target users assigned to A than B ($\alpha_\tau = 3$; the A oval is larger than the B oval), and the proportions of Poets and Quants who are targeted are equal ($\pi_\tau = 1$; the proportion of the blue and red quadrants inside the ovals are the same). But divergent delivery ($\rho_\tau = 8$) causes a higher proportion of the *targeted* A users to be Poets and a higher proportion of the *targeted* B users to be Quants ($\sigma_{\text{PA}} \neq \sigma_{\text{PB}}$; the vertical positioning of the center of the ovals is higher for A than B).

3.3 Estimation of effects and bias

The reason divergent delivery causes problems for causal inference in A/B comparisons is that varying targeting probabilities across user types and ads is equivalent to changing the posterior mixing probabilities among

Figure 4: Examples of How Ratios α_τ, π_τ, and ρ_τ Define Relationships among Targeting Probabilities $\Phi_{XZ} = \mathbf{P}(\tau = 1 \mid X, Z)$ and Posterior Mixtures $\sigma_{XZ} = \mathbf{P}(X \mid \tau = 1, Z)$ for Two Ads and Two User Types



Note to Figure 4: Each panel is an audience, divided into quadrants for each combination of ad (A on left, B on right), and user type (Poets in blue on the top, Quants in red on the bottom). As in Fig. 3a, Φ_{XZ} is the proportion of a quadrant inside a targeting oval, and σ_{XZ} is the proportion of a targeting oval that covers a quadrant. For example, the targeting probability Φ_{QA} is the proportion of the A-Quants who *are targeted* (the proportion of each bottom-left red ad-audience square inside the oval), and the posterior probability σ_{QA} is the proportion *of the targeted* A-users who are Quants (the proportion of left oval that is red). Each small grid square (white lines) represents 1% of the audience in a quadrant (e.g., if $\Phi_{PA} = .26$, the blue part of the A oval covers the equivalent of 26 squares). Panels are arranged by the ratios of marginal targeting probabilities between Poets and Quants (π_τ in each column), between ads A and B ($\alpha_\tau = 1$ in rows 1 and 3 and $\alpha_\tau = 3$ in rows 2 and 4)), and whether the platform engages in divergent delivery ("no" the top two $\rho_\tau = 1$ rows, and "yes" in the bottom two $\rho_\tau = 8$ rows). In all of these panels, $\gamma_P = \gamma_Q = .5$, $\zeta_A = \zeta_B = .5$, and $\widetilde{\Phi} = .1$.

the targeted set of users. In practice, the problem arises when experimental results reported to the advertiser, computed from outcomes of targeted users, do not actually reflect the effects the advertiser wants to measure.

We define $\lambda_{Z}^{\text{Targ}}$ as the lift among users assigned to, and *targeted* with, ad Z. Just like λ_{Z}^{ATE} (Eq. 7), $\lambda_{Z}^{\text{Targ}}$ is also a mixture of λ_{XZ}, but the targeted group's mixture weights are the *posterior* probabilities over types (σ_{XZ}), instead of the audience prior (γ_X).

$$\lambda_{Z}^{\text{Targ}} = \mathbf{E}\Big[Y_{Z}^{(1)} - Y_{Z}^{(0)} \,\Big|\, \tau = 1\Big] = \sum_{\forall X} \lambda_{XZ}\sigma_{XZ} \tag{16}$$

The A/B difference $\Delta_{AB}^{\text{Targ}}$ is a difference in lifts among users targeted with ads A and B, which is a difference-in-difference of expected potential outcomes.

$$\Delta_{AB}^{\text{Targ}} = \lambda_{A}^{\text{Targ}} - \lambda_{B}^{\text{Targ}} = \mathbf{E}\Big[Y_{A}^{(1)} - Y_{A}^{(0)} \,\Big|\, \tau = 1\Big] - \mathbf{E}\Big[Y_{B}^{(1)} - Y_{B}^{(0)} \,\Big|\, \tau = 1\Big] \tag{17}$$

$$= \sum_{\forall X} \lambda_{XA}\sigma_{XA} - \sum_{\forall X} \lambda_{XB}\sigma_{XB} = \big(\lambda_{x_1A}\sigma_{x_1A} - \lambda_{x_1B}\sigma_{x_1B}\big) + \big(\lambda_{x_2A}\sigma_{x_2A} - \lambda_{x_2B}\sigma_{x_2B}\big) + \cdots \tag{18}$$

Each parenthetical $\lambda_{xA}\sigma_{xA} - \lambda_{xB}\sigma_{xB}$ term in Eq. 18 is the contribution of type x to the A/B difference among the targeted users. The corresponding quantity among the audience is $\lambda_{xA}\gamma_x - \lambda_{xB}\gamma_x$ (Eq. 8). The difference between these targeted and audience quantities depends on the algorithm's targeting policies (α_τ, π_τ, ρ_τ), since only the mix changes from γ_X to σ_{XZ}. The λ_{XZ} are unaffected.

If targeting were entirely random ($\pi_\tau = 1, \rho_\tau = 1$), then $\sigma_{xA} = \sigma_{xB} = \gamma_x$. This is the only situation in which $\Delta_{AB}^{\text{Targ}} = \Delta_{AB}^{\text{ATE}}$. But if targeting is based only on user types, and not divergent across ads ($\pi_\tau \neq 1, \rho_\tau = 1$), then $\sigma_{xA} = \sigma_{xB} \neq \gamma_x$. Or if targeting by user type is divergent across ads ($\pi_\tau \neq 1, \rho_\tau \neq 1$), then $\sigma_{xA} \neq \sigma_{xB} \neq \gamma_x$. In those two non-random targeting cases, the expected difference between the lift of A among users targeted with A and the lift of B among users targeted with B (Eq. 17) is *not* equivalent to the difference between lifts of A and B when targeted to identical mixes by the overall campaign ($\rho_\tau = 1$).

3.3.1 Estimates

The distinctions between $\lambda_{Z}^{\text{Targ}}$ and λ_{Z}^{ATE}, and between $\Delta_{AB}^{\text{Targ}}$ and Δ_{AB}^{ATE}, are important because only the targeted values can be estimated from the data reported to the advertiser, while the advertiser may be interested in the effects on the audience. In a A/B/n test with holdout design (Sec. 2.2), the platform collects observed outcomes $Y_{Z}^{(\text{obs})} = Y_{Z}^{(1)}$ from users in the treatment arm ($\tau = 1, R = 1$), and $Y_{Z}^{(\text{obs})} = Y_{Z}^{(0)}$ from users in the holdout arm ($\tau = 1, R = 0$). The advertiser's report of experimental results contains only aggregated counts, sums, or averages of these observed results. Therefore, the *advertiser's estimate* of $\lambda_{Z}^{\text{Targ}}$ for each ad is the difference in sample means of observed outcomes for targeted users in the two arms of the test, and the estimate of $\Delta_{AB}^{\text{Targ}}$ is the difference in the estimates of those lifts.

$$\hat{\lambda}_{Z}^{\text{Targ}} = \bar{Y}_{Z,\text{Trt}}^{(1)} - \bar{Y}_{Z,\text{Hold}}^{(0)} \tag{19}$$

$$\widehat{\Delta}_{AB}^{\text{Targ}} = \hat{\lambda}_{A}^{\text{Targ}} - \hat{\lambda}_{B}^{\text{Targ}} \tag{20}$$

Equations 16 to 18 let us formalize the concerns from Sec. 2.2 that are facing advertisers who rely on the estimators in Eqs. 19 and 20 for inferences about the effectiveness of their ad creatives:

- The advertiser will see an ad's $\hat{\lambda}_Z^{\text{Targ}}$, which is an unbiased estimate for λ_Z^{Targ}, but will not get estimates of each λ_{XZ}. This aggregate estimate confounds changes to λ_{XZ} and σ_{XZ} because the user types X are unobserved. Multiplying λ_{XZ} by a constant, and dividing σ_{XZ} by that same constant, leaves λ_Z^{Targ} (and $\hat{\lambda}_Z^{\text{Targ}}$) unchanged.[15] Thus, the advertiser cannot know whether the observed lift from an ad is due to users responding positively to the ad's creative elements themselves, or to the algorithm's method of choosing which users will receive that ad.
- As long as assignment to treatment and holdout arms is random, the reported $\hat{\lambda}_Z^{\text{Targ}}$ is a valid ITT estimator of λ_Z^{ATET}. If the advertiser is comfortable limiting the scope of inference on lift to only targeted users of that one ad Z, then this estimator of λ_Z^{ATET} meets the advertiser's needs. But the advertiser who cares about inferring lift of an ad for the entire audience really does need an estimate of λ_Z^{ATE} instead. Because non-random targeting means that $\sigma_{XZ} \neq \gamma_X$, even the true values for λ_Z^{ATET} are not equal to λ_Z^{ATE}, and so estimating $\hat{\lambda}_Z^{\text{Targ}}$ does not help (Sec. 2.2, Concern 1).
- If the advertiser merely wants to compare outcomes between ads, regardless of the source of the differences, then $\widehat{\Delta}_{AB}^{\text{Targ}}$ satisfies those needs. But with divergent delivery, $\hat{\lambda}_A^{\text{Targ}}$ and $\hat{\lambda}_B^{\text{Targ}}$ are computed from different mixes of users ($\sigma_{XA} \neq \sigma_{XB}$). If the advertiser instead wants to separate the effect of ad creatives from how the targeting algorithm selects users for each ad, then the advertiser does not want $\widehat{\Delta}_{AB}^{\text{Targ}}$, because even the true values $\Delta_{AB}^{\text{ATET}}$ and $\Delta_{AB}^{\text{Targ}}$ do not equal Δ_{AB}^{ATE} (Sec. 2.2, Concern 2). An A/B/n test with holdout does not solve this problem.

3.3.2 Bias

For the rest of this paper, we will focus on how divergent delivery leads the estimated effect from the targeted users, $\widehat{\Delta}_{AB}^{\text{Targ}}$, to deviate from the effect for the audience, Δ_{AB}^{ATE}. We define another "difference-in-difference," $\hat{\mathcal{E}}_Z^{\lambda} = \hat{\lambda}_Z^{\text{Targ}} - \lambda_Z^{\text{ATE}}$, to be the *bias* in the estimate of an ad's lift computed from only targeted users, relative to the lift in the audience. Then, the difference between these values is the *bias* in the A/B difference.

$$\hat{\mathcal{E}}_{AB}^{\Delta} = \widehat{\Delta}_{AB}^{\text{Targ}} - \Delta_{AB}^{\text{ATE}} = \hat{\mathcal{E}}_A^{\lambda} - \hat{\mathcal{E}}_B^{\lambda} \tag{21}$$

Eq. 21, is a "diff-in-diff-in-diff" (DiDiD): first differences are between outcomes of treatment and holdout groups, $\hat{\lambda}_Z^{\text{Targ}} = \bar{Y}_{Z,\text{Trt}}^{(1)} - \bar{Y}_{Z,\text{Hold}}^{(0)}$, and the same for $\lambda_Z^{\text{ATE}} = \mathbf{E}\left[Y_Z^{(1)} - Y_Z^{(0)}\right]$; second differences are between ads $\widehat{\Delta}_{AB}^{\text{Targ}} = \hat{\lambda}_A^{\text{Targ}} - \hat{\lambda}_B^{\text{Targ}}$ and $\Delta_{AB}^{\text{ATE}} = \lambda_A^{\text{ATE}} - \lambda_B^{\text{ATE}}$; and third differences are between targeted and audience values, $\hat{\mathcal{E}}_{AB}^{\Delta} = \widehat{\Delta}_{AB}^{\text{Targ}} - \Delta_{AB}^{\text{ATE}}$. We use the term "bias" because $\hat{\mathcal{E}}_{AB}^{\Delta}$ is a difference between an estimate and "truth," with truth being the average treatment effect in the entire audience. This bias is not due to sampling or estimation error. In fact, from an A/B/n test with holdout experiment's data, the estimate of an ad's lift for its targeted users

[15] Formally, for any constant $c > 0$, $(c\lambda_{PZ})\,(\sigma_{PZ}/c) + \lambda_{QZ}\sigma_{QZ} = \lambda_{PZ}\sigma_{PZ} + \lambda_{QZ}\sigma_{QZ} = \lambda_Z^{\text{Targ}}$.

is estimated without problem. The remaining issue contributing to $\hat{\mathcal{E}}_{Z}^{\lambda}$ is that the targeted mix and audience mix differ: $\hat{\lambda}_{Z}^{\text{Targ}} \neq \hat{\lambda}_{Z}^{\text{ATE}}$. But that bias alone for a *single ad* is not our focal concern. Instead, we focus on the way that those targeted-vs-audience gaps differ across ads in Eq. 21.

To formally study the factors that are causing this bias, we continue with the case of two ads, *A* and *B*, and two user types, Poets and Quants. The progression of Eqs. 22 to 25 comprises a derivation of $\mathcal{E}_{AB}^{\Delta}$ for this 2×2 case. Equations 22a to 22f collect and simplify the general expressions for ad-specific lift in Eqs. 7 and 16. Equations 23 to 25 follow, as special cases of Eqs. 8, 18 and 21.

$$\lambda_{A}^{\text{ATE}} = \gamma_{P}\lambda_{PA} + (1-\gamma_{P})\,\lambda_{QA} \tag{22a}$$

$$\lambda_{A}^{\text{Targ}} = \sigma_{PA}\lambda_{PA} + (1-\sigma_{PA})\,\lambda_{QA} \tag{22b}$$

$$\hat{\mathcal{E}}_{A}^{\lambda} = (\sigma_{PA}-\gamma_{P})\,(\lambda_{PA}-\lambda_{QA}) \tag{22c}$$

$$\lambda_{B}^{\text{ATE}} = \gamma_{P}\lambda_{PB} + (1-\gamma_{P})\,\lambda_{QB} \tag{22d}$$

$$\lambda_{B}^{\text{Targ}} = \sigma_{PB}\lambda_{PB} + (1-\sigma_{PB})\,\lambda_{QB} \tag{22e}$$

$$\hat{\mathcal{E}}_{B}^{\lambda} = (\sigma_{PB}-\gamma_{P})\,(\lambda_{PB}-\lambda_{QB}) \tag{22f}$$

$$\Delta_{AB}^{\text{ATE}} = \gamma_{P}\,(\lambda_{PA}-\lambda_{PB}) + (1-\gamma_{P})\,(\lambda_{QA}-\lambda_{QB}) \tag{23}$$

$$\Delta_{AB}^{\text{Targ}} = \left(\sigma_{PA}\lambda_{PA} + (1-\sigma_{PA})\,\lambda_{QA}\right) - \left(\sigma_{PB}\lambda_{PB} + (1-\sigma_{PB})\,\lambda_{QB}\right) \tag{24}$$

$$\mathcal{E}_{AB}^{\Delta} = \underbrace{\overbrace{(\sigma_{PA}-\gamma_{P})}^{(25.A.1)}\;\overbrace{(\lambda_{PA}-\lambda_{QA})}^{(25.A.2)}}_{(25.A)} - \underbrace{\overbrace{(\sigma_{PB}-\gamma_{P})}^{(25.B.1)}\;\overbrace{(\lambda_{PB}-\lambda_{QB})}^{(25.B.2)}}_{(25.B)} \tag{25}$$

Equation 25 shows that bias comes from three sources:

- *The targeted mix differs from the audience.* For each ad, Factors 25.A.1 and 25.B.1 quantify how much the mixture of types among targeted users differs from the mixture of types in the audience. The bias is smaller when the proportion each type among targeted users is similar to the proportion in the audience.
- *Users respond differently to the same ad.* Factors Eq. 25.A.2 and 25.B.2 quantify the differences between the lifts for users with each latent type. The bias is smaller when the different user types are more homogeneous.
- *The targeted mix of one ad differs from the targeted mix of the other ad.* Terms Eq. 25.A and Eq. 25.B show how heterogeneity of users' responsiveness to ads moderates the size and magnitude of a targeting policy's effect on the bias.

In Eq. 25, the amount of the bias depends on heterogeneity in users' responses to the two ads through the $(\lambda_{PA}-\lambda_{QA})$ and $(\lambda_{QB}-\lambda_{PB})$ factors, which are the partial derivatives of $\mathcal{E}_{AB}^{\Delta}$ with respect to the targeted mix of users. For example, if *A*-Poets are more responsive than *A*-Quants, then targeting more Poets among the *A* users will push $\hat{\lambda}_{A}^{\text{Targ}}$ above λ_{A}^{ATE}, driving up $\widehat{\Delta}_{AB}^{\text{Targ}}$ relative to Δ_{AB}^{ATE}, and increasing the bias. But if *B*-Poets are more responsive than *B*-Quants as well, then targeting more Poets among the *B*-users makes $\hat{\lambda}_{B}^{\text{Targ}}$ larger than λ_{B}^{ATE}, driving *down* $\widehat{\Delta}_{AB}^{\text{Targ}}$ and offsetting the increase in bias from *A*. If *B*-Poets were less responsive than *B*-Quants, then targeting ad *B* would create even more bias. Without some structure in how we express the relationships

among λ_{XZ}, it can all get quite confusing, especially since *the advertiser does not observe lifts for each user type separately*. We discuss certain aspects and properties of these marginal effects in the Appendix.

3.4 Characterizing response heterogeneity in an audience

To simplify and structure how the responsiveness of user types to ads, we express those relationships in terms of main effects and an interaction among potential outcomes. To help with notation, define $\Theta^{(1)}_{XZ} = \mathbf{E}\big[\, Y^{(1)}_{Z} \mid X \,\big]$, $\Theta^{(0)}_{XZ} = \mathbf{E}\big[\, Y^{(0)}_{Z} \mid X \,\big]$, $\Theta^{(1)}_{Z} = \mathbf{E}\big[Y^{(1)}_{Z}\big]$, $\Theta^{(0)}_{Z} = \mathbf{E}\big[Y^{(0)}_{Z}\big]$, $\Theta^{(1)}_{X} = \mathbf{E}\big[\, Y^{(1)} \mid X = \mathrm{X} \,\big]$, and $\Theta^{(0)}_{Z} = \mathbf{E}\big[\, Y^{(0)} \mid X = \mathrm{Z} \,\big]$ (Sec. 3.1). Thus, we can write Eq. 6 as $\lambda_{XZ} = \Theta^{(1)}_{XZ} - \Theta^{(0)}_{XZ}$ and Eq. 7 as $\lambda^{\text{ATE}}_{Z} = \Theta^{(1)}_{Z} - \Theta^{(0)}_{Z}$. Using the same logic behind definitions of ratios of targeting probabilities (Eqs. 13 to 15, with subscript τ), we summarize the pairwise relationships among $\Theta^{(1)}_{X}$, $\Theta^{(1)}_{Z}$, and $\Theta^{(1)}_{XZ}$ with ratios of expected potential outcomes between ads (α_{Y}) , between user types (π_{Y}), and an interaction between user type and ad (ρ_{Y}).

$$\alpha_{Y} = \frac{\Theta^{(1)}_{A}}{\Theta^{(1)}_{B}} = \frac{\gamma_{P}\Theta^{(1)}_{PA} + \gamma_{Q}\Theta^{(1)}_{QA} + \cdots}{\gamma_{P}\Theta^{(1)}_{PB} + \gamma_{Q}\Theta^{(1)}_{QB} + \cdots} \qquad \text{(ad effectiveness)} \tag{26}$$

$$\pi_{Y} = \frac{\Theta^{(1)}_{P}}{\Theta^{(1)}_{Q}} = \frac{\zeta_{A}\Theta^{(1)}_{PA} + \zeta_{B}\Theta^{(1)}_{PB} + \cdots}{\zeta_{A}\Theta^{(1)}_{QA} + \zeta_{B}\Theta^{(1)}_{QB} + \cdots} \qquad \text{(user heterogeneity)} \tag{27}$$

$$\rho_{Y} = \frac{\Theta^{(1)}_{PA}}{\Theta^{(1)}_{PB}} \Big/ \frac{\Theta^{(1)}_{QA}}{\Theta^{(1)}_{QB}} \qquad \text{(user-ad response interaction)} \tag{28}$$

The α_{Y} ratio captures relative *ad effectiveness* overall. If $\alpha_{Y} > 1$, the expected response after exposure to ad *A* is greater than ad *B*. The π_{Y} and ρ_{Y} ratios describe *response heterogeneity*. The π_{Y} ratio alone denotes *user heterogeneity* overall. If $\pi_{Y} > 1$ the expected response from exposed Poets is higher than from exposed Quants, on average. The odds ratio ρ_{Y} operationalizes a *user-ad response interaction*. If $\rho_{Y} > 1$, Poets respond even better to *A* and Quants respond even better to *B* than whatever the marginal ratios α_{Y} and π_{Y} alone would have dictated. Because these ratios define the exact same relationships as for targeting policies, we refer the reader back to Fig. 4 for a visualization.

4 Bringing targeting and heterogeneity together

We now tie together the various pieces of our framework:

- α_{τ}, π_{τ}, and ρ_{τ} characterize *targeting policies* as relationships among targeting probabilities Φ_{XZ} for all X and Z, and equivalently, conditional proportions of types among users targeted with each ad, σ_{PA} and σ_{PB}.
- α_{Y}, π_{Y}, and ρ_{Y} characterize *response heterogeneity* in users' potential outcomes after being exposed to ads, and consequently, in the potential lifts of each ad for each user type, λ_{XZ} for all X and Z.
- The estimated lift for each ad, $\hat{\lambda}^{\text{Targ}}_{Z}$ is a mix of λ_{XZ}, with σ_{PZ} providing the weights for the mixture, and $\widehat{\Delta}^{\text{Targ}}_{AB}$ is the difference in those estimated lifts.

- The bias in A/B difference, $\mathcal{E}^{\Delta}_{AB}$, is the difference between ads and between the true value for audience Δ^{ATE}_{AB} and estimate for targeted users Δ^{Targ}_{AB}.

4.1 The relationship between the mix of targeted users and estimated aggregate lifts

Figure 5 illustrates how targeting and heterogeneity work together to govern how estimates computed from targeted users deviate from true values in the audience. The annotations in the figure walk through how the mix of user types determines the ad's aggregate lift. Each ad's targeted lift λ^{Targ}_{Z} (y-axis; Eqs. 22b and 22e) is a linear combination of the type-specific lifts, λ_{PZ} and λ_{QZ} (endpoints of the diagonal lines), weighted by the proportion of targeted Quants relative to Poets, σ_{QZ} and $1-\sigma_{QZ}$ (x-axis). If the algorithm were to target an ad to only one user type (e.g., all Poets), the aggregate lift would equal type-specific lift (e.g., λ_{PZ} at the endpoint). And if the algorithm were to target ads randomly to the same targeted mix equal to the audience mix for both ads (i.e., $\sigma_{PZ} = \sigma_{QZ} = \gamma_{Q}$), then aggregate lifts would be λ^{ATE}_{Z} (green lines and notes).

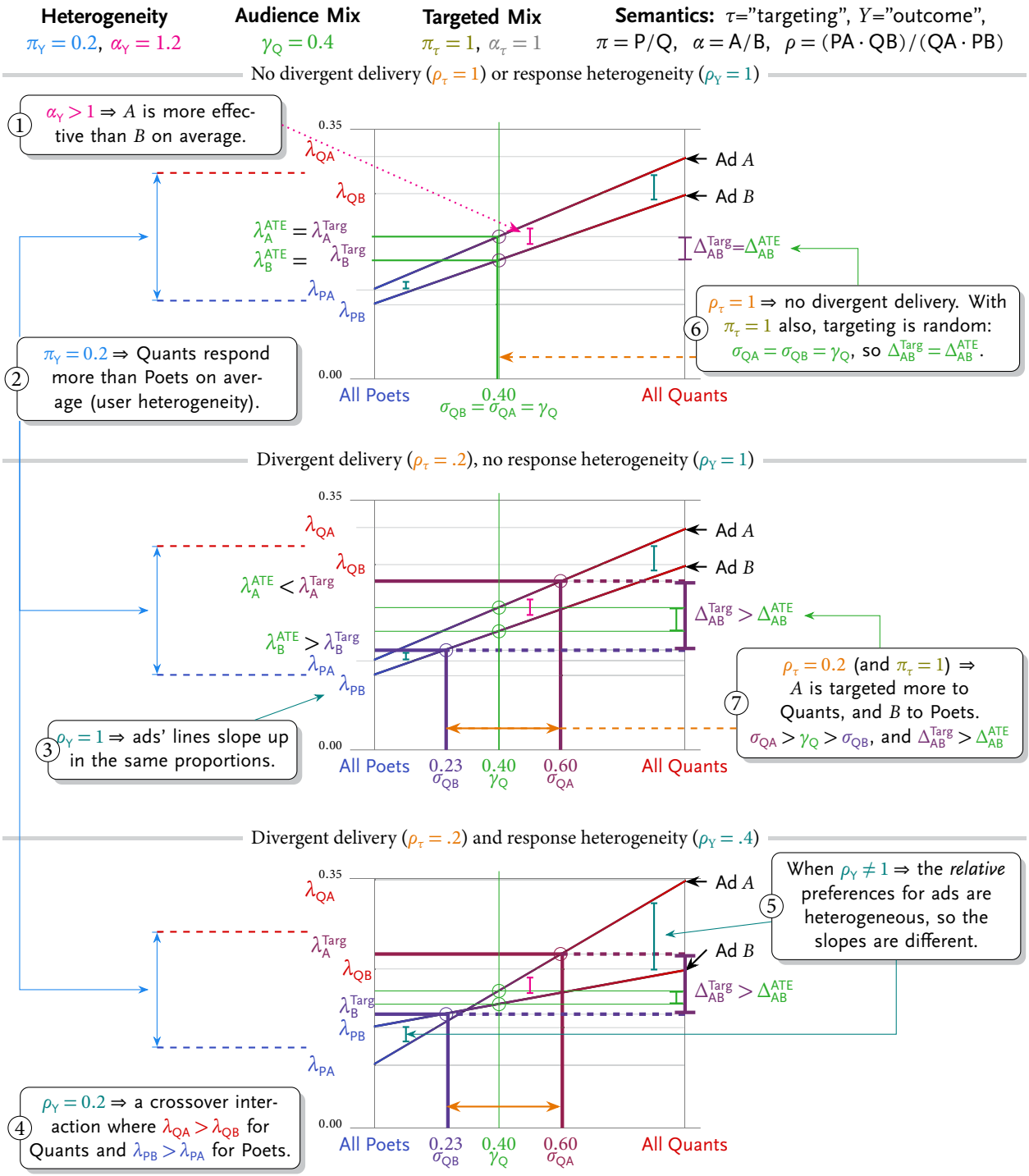
Panels in Fig. 5 are distinguished by divergent delivery (ρ_{τ}) and the user-ad response interaction (ρ_{Y}). With divergent delivery ($\rho_{\tau} = .2$, middle and bottom panels), the targeted mixes are different for each ad, so σ_{QA} and σ_{QB} deviate from γ_{Q}, and from each other (horizontal separation between σ_{QA} and σ_{QB}). When the mix changes in favor of the better responding user type for an ad (e.g., σ_{QA} increases), the estimated aggregate lift of the targeted mix for that ad increases above its true lift in the audience (e.g., $\lambda^{Targ}_{A} > \lambda^{ATE}_{A}$). The rate of that increase in aggregate lift with respect to change in the mix (the slope of the line) depends on the heterogeneity in lifts between user groups for each ad (i.e., vertical separation of the endpoints λ_{PZ} and λ_{QZ}). The *same level of divergent delivery* has different effects depending on the response heterogeneity. With no response heterogeneity ($\rho_{Y} = 1$, middle panel), $\lambda_{XA} > \lambda_{XB}$ for all users (non-intersecting lines). As the mix favors the Quants (the stronger responders), the stronger ad's lift is overestimated and the weaker ad's is underestimated, thereby overestimating the difference Δ^{Targ}_{AB} relative to Δ^{ATE}_{AB}. In contrast, when $\rho_{Y} = .4$ (bottom panel), $\lambda_{QA} > \lambda_{QB}$, but $\lambda_{PB} > \lambda_{PA}$ (a crossover interaction). The size of the bias is reduced for ad *B*'s lift but increased substantially for ad *A*'s lift, which generates a larger bias in their differences.

4.1.1 Simpson's Reversal

An extreme example where the estimated $\widehat{\Delta}^{Targ}_{AB}$ is a poor estimate of the true Δ^{ATE}_{AB} is when the *signs* of the two effects are different. This is an example of an undetectable *Simpson's reversal,* a pattern of aggregation bias across heterogeneous groups (Simpson 1951; Blyth 1972; Baker and Kramer 2001; Pearl 2014). A Simpson's reversal occurs when the true lift of *A* is greater than *B* for each user type separately, but the estimates of lift when aggregated across unobserved user types show that ad *B* is stronger than *A*; that is, if $\lambda_{PA} > \lambda_{PB}$ and $\lambda_{QA} > \lambda_{QB}$, but $\lambda^{Targ}_{A} < \lambda^{Targ}_{B}$.

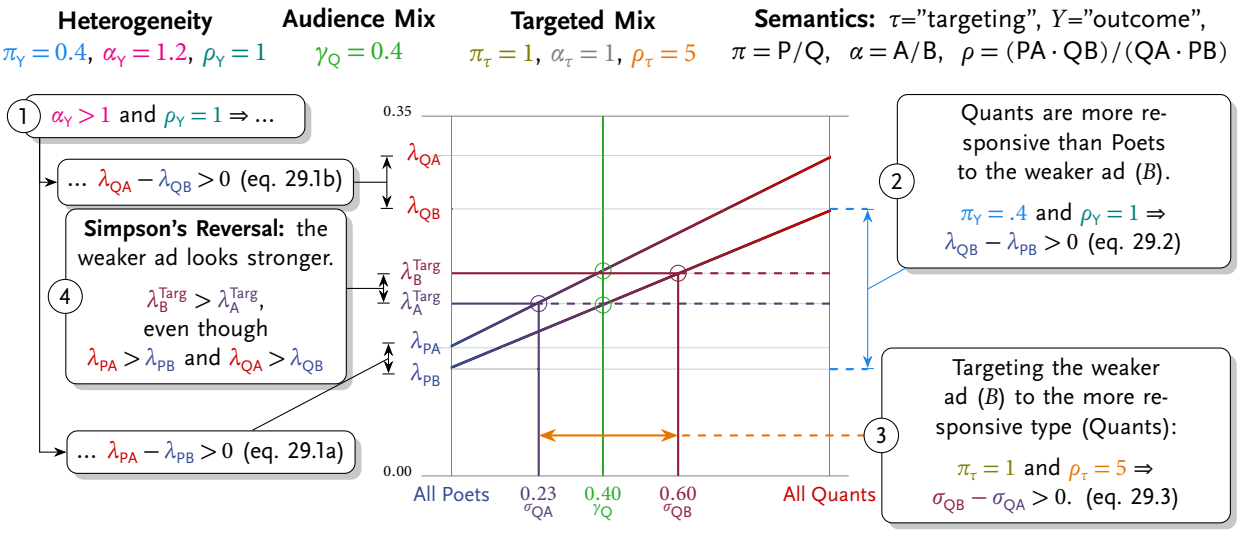
Figure 6 illustrates how a Simpson's reversal can happen. In this example, *A* has a higher lift than *B* for both user types, but Quants are overall so much more responsive than Poets that any ad's mixture in which Quants are

Figure 5: Effects of Targeting and Heterogeneity on Aggregate Lift



Note to Figure 5: Targeted aggregate lifts λ_Z^{Targ} (y-axis) are linear combinations of λ_{PZ} and λ_{QZ} weighted by targeted (posterior) mixture probabilities σ_{QA} and σ_{QB} (x-axis). These weights deviate from audience (prior) mixture probability γ_Q. In all panels, $\alpha_Y = 1.2$ (on average, A is more effective than B), $\pi_Y = 0.2$ (on average, Quants respond more than Poets), $\gamma_Q = .40$ (40% of the audience are Quants), $\bar{\Phi} = .2$ (the overall targeting probability), $\pi_\tau = 1$ (on average, Quants and Poets are targeted equally), and $\zeta_A = .5$ (balanced random assignment to ads).

Figure 6: Visualizing the Simpson's Reversal Conditions from Eq. 29



overrepresented will make that ad appear to be stronger than it actually is in the audience. A targeting policy with $\rho_\tau = 5$ creates enough divergent delivery that the estimated $\hat{\lambda}_B^{Targ}$ will be too high, the estimated $\hat{\lambda}_A^{Targ}$ will be too low, and a Simpson's reversal will occur.

Mathematically, a Simpson's reversal will occur when the following inequality holds.

$$\overbrace{(\lambda_{QB} - \lambda_{PB})}^{(29.2)} \overbrace{(\sigma_{QB} - \sigma_{QA})}^{(29.3)} > \underbrace{\overbrace{(1 - \sigma_{QA})}^{} \overbrace{(\lambda_{PA} - \lambda_{PB})}^{(29.1a)} + \sigma_{QA} \overbrace{(\lambda_{QA} - \lambda_{QB})}^{(29.1b)}}_{(29.1)} \tag{29}$$

Equation 29 holds when: (1) the amount by which the *stronger ad's* lift exceeds the weaker ad's lift among targeted users *within each user type* is sufficiently small (29.1); (2) the difference between user types for the *weaker* ad's lift is sufficiently large (29.2); and (3) the users responding better to the weaker ad are more prevalent among users targeted with that weaker ad than among users targeted with the stronger ad (29.3). When A is the stronger ad (so 29.1 is positive), these conditions will hold when ρ_Y and α_Y are close enough to 1, π_Y and ρ_τ are far enough from 1, and ρ_τ and π_Y are on opposite sides of 1.

We can address some common questions about when and why these conditions for Simpson's reversal may arise, and how an advertiser might detect it.

- **How will an advertiser know if $\hat{\Delta}_{AB}^{Targ}$ reflects a Simpson's reversal?** They won't. The advertiser observes the estimated aggregate lifts $\hat{\lambda}_A^{Targ}$ and $\hat{\lambda}_B^{Targ}$, but none of the true, type-specific lifts, λ_{XZ}. The Simpson's reversal would be undetectable, so *the advertiser will not know if the sign of the A/B test is different from the effect they are trying to learn.*
- **How common might an unobserved Simpson's reversal be?** To effectively reach a heterogeneous mix of users, advertisers and platforms want to exploit differences in predicted responses to ads. If ads in a campaign

share some common creative elements (e.g., reflect a common positioning strategy), then it is plausible for the difference in the lifts of those ads to be small ($\alpha_Y \approx 1$) and for unobserved heterogeneity in responses to be similar for both ads ($\rho_Y \approx 1$).[16] In that case, advertisers should prefer the algorithm to be cautious about changing the mix of types targeted with each ad by too much. But if the algorithm overestimates the differences among ads, then the targeting decisions may be more extreme than the true response heterogeneity might warrant. In that case, σ_{PA} and σ_{PB} could separate enough to create a Simpson's reversal.

- **Why would the algorithm even try to target ad *B* to Quants when *A* performs better among Quants?** Because the advertiser is conducting an experiment! In a non-experimental campaign, a targeting algorithm that suspects *A* will be stronger among both types might only target users who were assigned to *A*, and none who were assigned to *B*. But an experiment to compare *A* and *B* needs to expose at least some users to *B*, even though it is the weaker ad overall. Given the high degree of heterogeneity between types, *B*-Quants will still outperform *A*-Poets, so in an experiment the algorithm might aim to get as many conversions as it can from *B* out of the Quants. *The requirements of the experimental design could force (or at least lightly nudge) the algorithm toward targeting policies that make a Simpson's reversal more likely.*

In Sec. 4.2.2 (Example 5), we will present simulation results that reflect a Simpson's reversal when the algorithm *overtargets* based on small differences between ads.

4.2 Simulation

Next, we demonstrate through numerical simulation how divergent delivery and response heterogeneity conspire to cause a gap between $\widehat{\Delta}_{AB}^{Targ}$ and Δ_{AB}^{ATE} ($\mathcal{E}_{AB}^{\Delta} \neq 0$). The simulation will reveal how those effects are moderated by conditions described by the sets of ratios, α_Y, π_Y, ρ_Y and $\alpha_\tau, \pi_\tau, \rho_\tau$. And in some cases, the reported A/B difference even reverses the sign of the true effect for the audience.

4.2.1 Simulation assumptions and definitions

We provide the finer details of the simulation in the Web Appendix, and focus on the most important aspects here. The unit of analysis is a simulated "ad-audience dyad." For the purpose of the simulation, we will refer to averages of replicates of dyads with the same parameters simply as an "audience." The audience includes the users characterized by both the relative responsiveness of user types to ads (described by α_Y, π_Y, and ρ_Y), and the targeting policies applied to the users in that audience (described by α_τ, π_τ, and ρ_τ). The audience consists of two types of users, $X \in \{P, Q\}$ (which we continue to call Poets and Quants), proportioned equally ($\gamma_P = \gamma_Q = 1/2$). Each experiment is a A/B/n test with holdout with three ads $Z \in \{A, B, C\}$, to which users in the audience are

[16] We assume that the upper bound for our between-ad (A-vs-B) effect sizes will still be small, roughly on the order of single-ad (ad-vs-no-ad) effect sizes shown in large meta-analyses, such as, in online advertising (Johnson et al. 2017b), online social media advertising (Gordon et al. 2019), and television advertising (Lodish et al. 1995; Shapiro et al. 2020).

Table 2: Definitions of Simulated Quantities.

	"True" audience	Estimated	Bias
Lift for $Z \in \{A, B\}$	$\lambda_{Z}^{\text{ATE}} = \bar{Y}_{Z,\text{All}}^{(1)} - \bar{Y}_{Z,\text{All}}^{(0)}$	$\hat{\lambda}_{Z}^{\text{Targ}} = \bar{Y}_{Z,\text{Trt}}^{(1)} - \bar{Y}_{Z,\text{Hold}}^{(0)}$	$\hat{\mathcal{E}}_{Z}^{\lambda} = \hat{\lambda}_{Z}^{\text{Targ}} - \lambda_{Z}^{\text{ATE}}$
A/B difference	$\Delta_{AB}^{\text{ATE}} = \lambda_{A}^{\text{ATE}} - \lambda_{B}^{\text{ATE}}$	$\widehat{\Delta}_{AB}^{\text{Targ}} = \hat{\lambda}_{A}^{\text{Targ}} - \hat{\lambda}_{B}^{\text{Targ}}$	$\hat{\mathcal{E}}_{AB}^{\Delta} = \hat{\mathcal{E}}_{A}^{\lambda} - \hat{\mathcal{E}}_{B}^{\lambda} = \widehat{\Delta}_{AB}^{\text{Targ}} - \Delta_{AB}^{\text{ATE}}$

randomly assigned with equal probabilities ($\zeta_A = \zeta_B = \zeta_C = 1/3$). Additionally, the simulation invokes the following assumptions:

- An outcome is akin to a "conversion", meaning that all Y_i are binary random variables.
- Expected potential outcomes are conversion probabilities, conditional on being exposed or unexposed: $\Theta_{XZ}^{(0)} = \mathbf{P}\left(Y_Z^{(0)} = 1 \mid X\right)$ and $\Theta_{XZ}^{(1)} = \mathbf{P}\left(Y_Z^{(1)} = 1 \mid X\right)$. The expected outcomes for *unexposed* users vary by user type, but not the ad to which they were initially assigned. That is, $\Theta_{XA}^{(0)} = \Theta_{XB}^{(0)} = \Theta_{XC}^{(0)} = \Theta_{X}^{(0)}$ for each X.
- The average probability of any user being targeted with any ad is $\widetilde{\Phi} = \mathbf{P}(\tau = 1) = .2$, acting as a budget constraint. The average conversion probability for all users, across all ads, is $\widetilde{\Theta}^{(1)} = \mathbf{P}\left(Y^{(1)} = 1\right) = .2$.
- $V_i = 1$ for all users, and $\mathbf{P}(R = 1)$ is same for all users.

The "experimental conditions" of the simulation are selections from $\rho_Y \in \{1/8, 1, 8\}$ and $\rho_\tau \in \{1/8, 1/3, 1, 3, 8\}$, and simulated values of α_Y, π_Y, α_τ, and π_τ. Conditional on those parameters, we sample or solve for all Φ_{XZ}, $\Theta_{XZ}^{(0)}$, and $\Theta_{XZ}^{(0)}$, and generate the complete set of $2n_Z$ potential outcomes for each user (see the Web Appendix). We then follow the process tree in Fig. 2 to randomly assign users to ads, target users to those ads, and to generate the advertiser's "observed data." Proportions $\bar{Y}_{Z,\text{Trt}}^{(1)}$ and $\bar{Y}_{Z,\text{Trt}}^{(0)}$ are computed by tallying corresponding potential outcomes of users who were targeted and are in the treatment arm, while $\bar{Y}_{Z,\text{Hold}}^{(1)}$ and $\bar{Y}_{Z,\text{Hold}}^{(0)}$ are computed from targeted users in holdout arm. Because we have simulated *all* of the potential outcomes for *all* users, we can infer counterfactual effects for the entire audience, regardless of users' ad assignment or exposure status. The "true" effects among all simulated users in an audience are λ_A^{ATE}, λ_B^{ATE}, and Δ_{AB}^{ATE}, while $\hat{\lambda}_B^{\text{Targ}}$, $\hat{\lambda}_B^{\text{Targ}}$, and $\widehat{\Delta}_{AB}^{\text{Targ}}$ mimic the estimated effects that the platform would compute on behalf of the advertiser. Simulated values of $\hat{\mathcal{E}}_A^{\lambda}$, $\hat{\mathcal{E}}_B^{\lambda}$, and $\hat{\mathcal{E}}_{AB}^{\Delta}$ follow. Table 2 defines effects and biases in terms of these simulated proportions.

4.2.2 Simulation results

Next, we examine how heterogeneity and targeting policies interact to lead to different forms of estimation gaps. In Figs. 7 and 8 the y-axes respectively show $\hat{\mathcal{E}}_A^{\lambda}$ and $\hat{\mathcal{E}}_B^{\lambda}$ in the top and middle rows of panels, and $\hat{\mathcal{E}}_{AB}^{\Delta}$ in the bottom rows. The logic of any one of these panels is as follows. The six ratios define the relationships among the targeting probabilities and among the user response propensities. Those govern the mix of users targeted with each ad. Aggregating across each mix, we get ad-specific lift estimates and the estimated A/B difference. As these

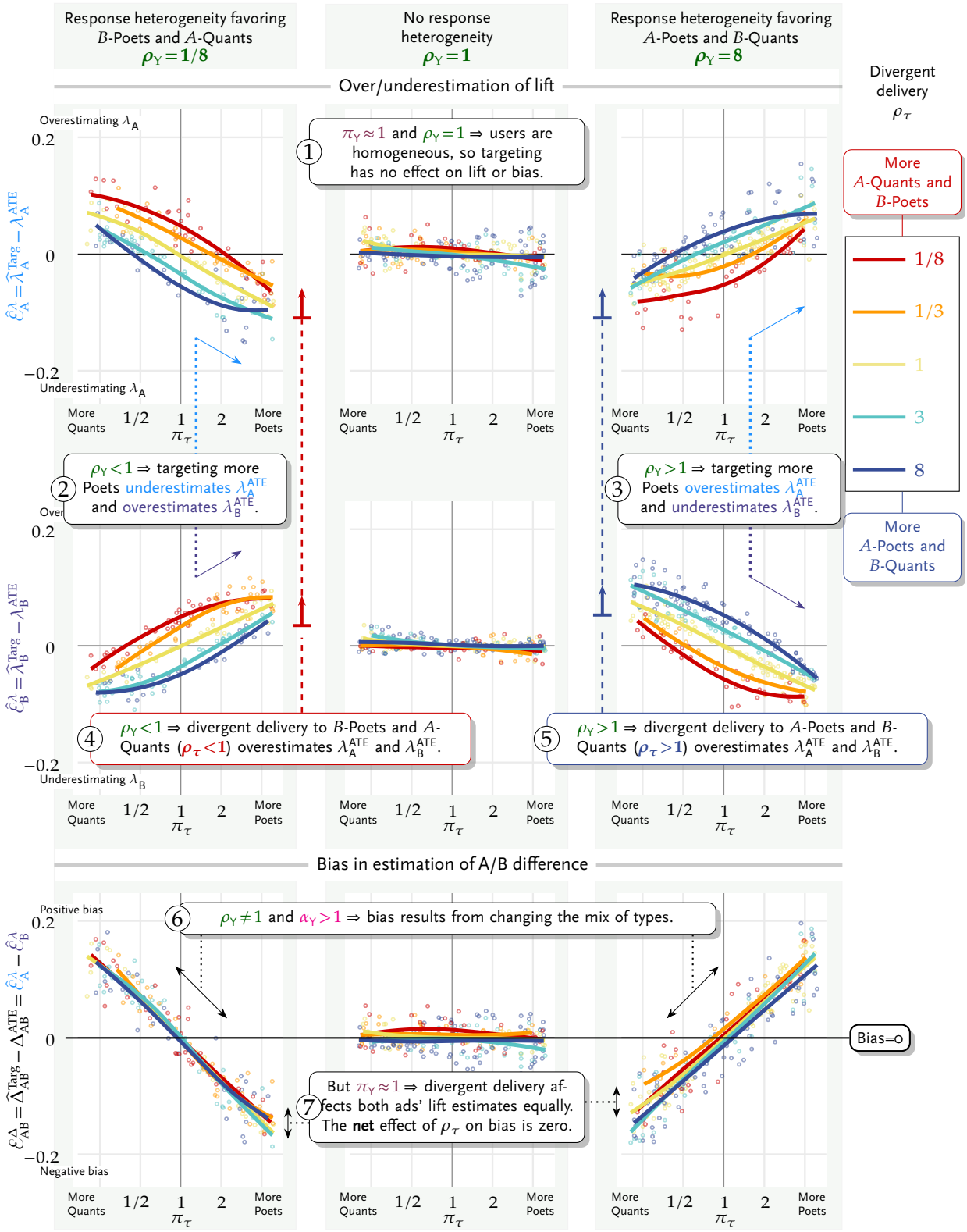
mixes differ from each other and from the audience mix, those estimates deviate from the true effects of ads on the audience. The nature of that deviation is driven by targeting and is moderated by heterogeneity.

- **Panel columns in Figs. 7 and 8 differ by parameters that govern user responses.** Audiences (circles) are classified into "worlds" according to their discretized ad responsiveness parameters which vary by column (ρ_Y) and by figure (π_Y). Each column of panels corresponds to a degree and direction of response heterogeneity, $\rho_Y \in \{1/8, 1, 8\}$. The user-ad response interaction dictates the degree to which the *A*-Poets and *B*-Quants have higher lift propensities ($\rho_Y = 8$, right column) or the *A*-Quants and *B*-Poets have the higher lift propensities ($\rho_Y = 1/8$, left column). And with no user-ad response interaction, the ratio of response propensities of Poets to Quants is the same for users assigned to each ad ($\rho_Y = 1$, middle column). For all panels in Figs. 7 and 8, ad *A* is stronger than *B* ($\alpha_Y > 1$), in aggregate across user types. For the audiences in Fig. 7, Poets respond about as much as Quants overall ($2/3 < \pi_Y < 4/3$, which we abbreviate as $\pi_Y \approx 1$), while in Fig. 8, Poets respond more than Quants ($\pi_Y > 4/3$).
- **Within panels in Figs. 7 and 8, parameters governing targeting decisions differ.** The targeting algorithm's policies differ for audiences within a panel in two ways. First, the type-specific targeting (π_τ, continuous on the *x*-axis) describes mixes ranging from more Quants ($\pi_\tau < 1$) to more Poets ($\pi_\tau > 1$), averaged across all ads. Second, divergent delivery (ρ_τ on the discrete color scale) favors a range of mixes from more *A*-Quants and *B*-Poets ($\rho_\tau = 1/8$; red) to more *A*-Poets and *B*-Quants ($\rho_\tau = 8$; blue). In the case of no divergent delivery ($\rho_\tau = 1$; yellow), the mix of targeted Poets to Quants is the same for each ad. In the simulation, α_τ varies in a tight interval around 1.

Rather than go through all possible combinations of effects in Figs. 7 and 8, we explore three columns (Examples 1, 2, and 3 below) to highlight the simulated audiences with the most extreme effects, and show how different parameter values attenuate those effects.

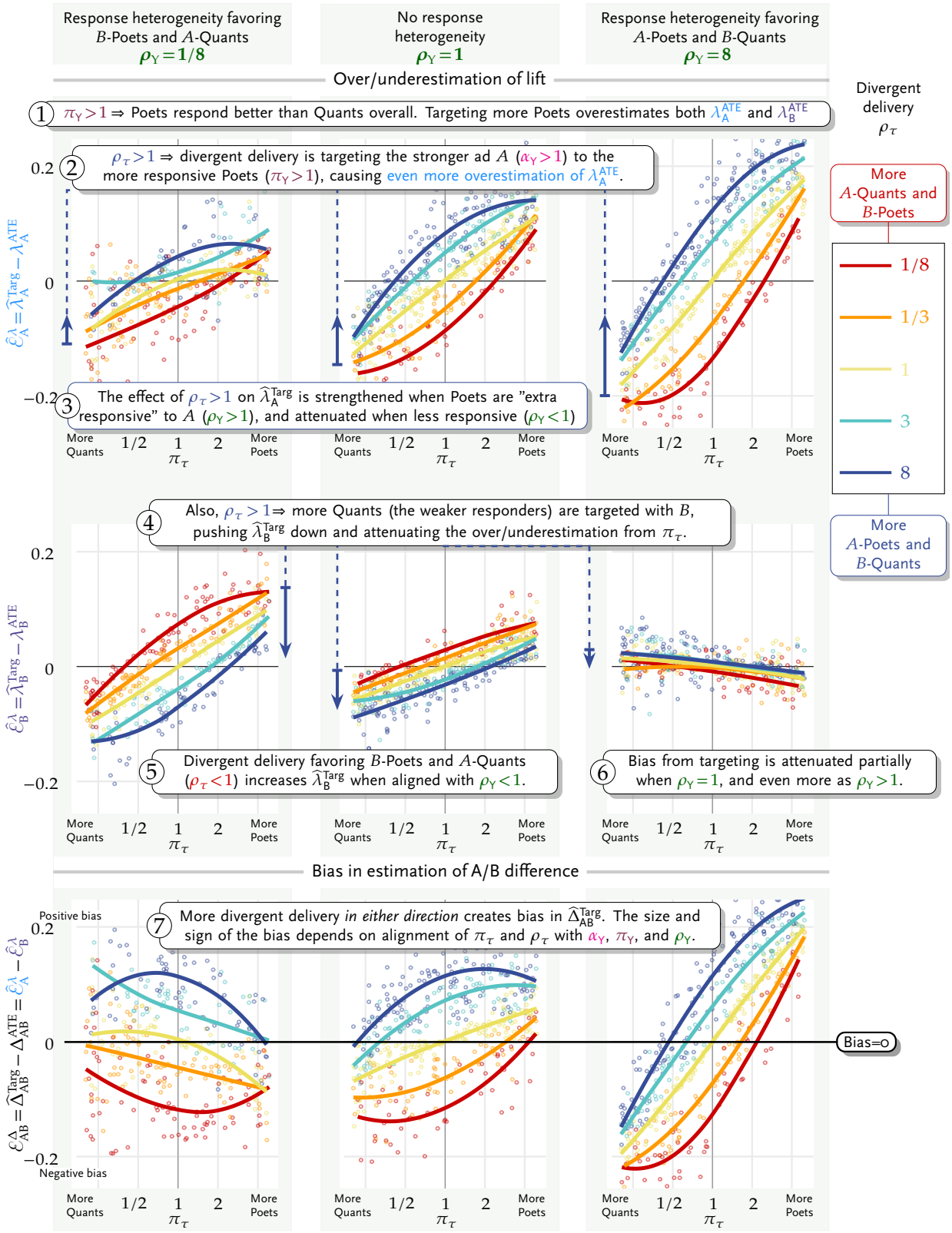
Example 1 (Fig. 7, left column; $\alpha_Y > 1$, $\pi_Y \approx 1$, $\rho_Y = 1/8$) For all panels of Fig. 7, ad *A* is stronger than ad *B*, overall ($\alpha_Y > 1$), and Poets and Quants respond similarly on average ($\pi_Y \approx 1$). But users still differ: audiences have extreme user-ad interaction ($\rho_Y = 1/8$), where *A*-Quants and *B*-Poets have higher lift propensities than their marginal effects α_Y and π_Y would suggest. In this left column, the top panel describes the bias in lift for ad *A* ($\hat{\mathcal{E}}^{\lambda}_{A}$). Changing the mix of the targeted users through π_τ (*x*-axis) and ρ_τ (color) affects the gap between $\hat{\lambda}^{\text{Targ}}_{A}$ and λ^{ATE}_{A}. We start by considering when the algorithm targets more Quants than Poets ($\pi_\tau < 1$, left side of *x*-axis; e.g., $\pi_\tau = 1/4$ implies a 1 : 4 ratio of Poets to Quants), compared to the audience mix ($\gamma_P = 1/2$ implies a 1 : 1 ratio). When the algorithm also employs divergent delivery favoring *A*-Quants and *B*-Poets ($\rho_\tau = 1/8$; red line), the resulting mix of users targeted with *A* skews more to the ad's best responders, *A*-Quants over *A*-Poets, than in the mix in the audience. This pushes the estimate of the lift for the targeted mix ($\hat{\lambda}^{\text{Targ}}_{A}$) above the true lift for the audience (λ^{ATE}_{A}), so the aggregate lift of ad *A* is overestimated ($\hat{\mathcal{E}}^{\lambda}_{A} > 0$).

Figure 7: Simulated $\hat{\mathcal{E}}^{\lambda}_{A}$, $\hat{\mathcal{E}}^{\lambda}_{B}$, and $\hat{\mathcal{E}}^{\Delta}_{AB}$ when $A \succ B$ ($\alpha_Y > 1$) and $\Theta^{(1)}_P \approx \Theta^{(1)}_Q$ ($\pi_Y \approx 1$).



Note to Figure 7: Follow the numbered signposts.

Figure 8: Simulated $\hat{\mathcal{E}}^{\lambda}_{A}$, $\hat{\mathcal{E}}^{\lambda}_{B}$, and $\hat{\mathcal{E}}^{\Delta}_{AB}$ when $A \succ B$ ($\alpha_Y > 1$) and $\Theta^{(1)}_P > \Theta^{(1)}_Q$ ($\pi_Y > 1$).



Note to Figure 8: Follow the numbered signposts. Fig. 7 ($\pi_Y \approx 1$) and Fig. 8 ($\pi_Y > 1$) are distinguished by user heterogeneity.

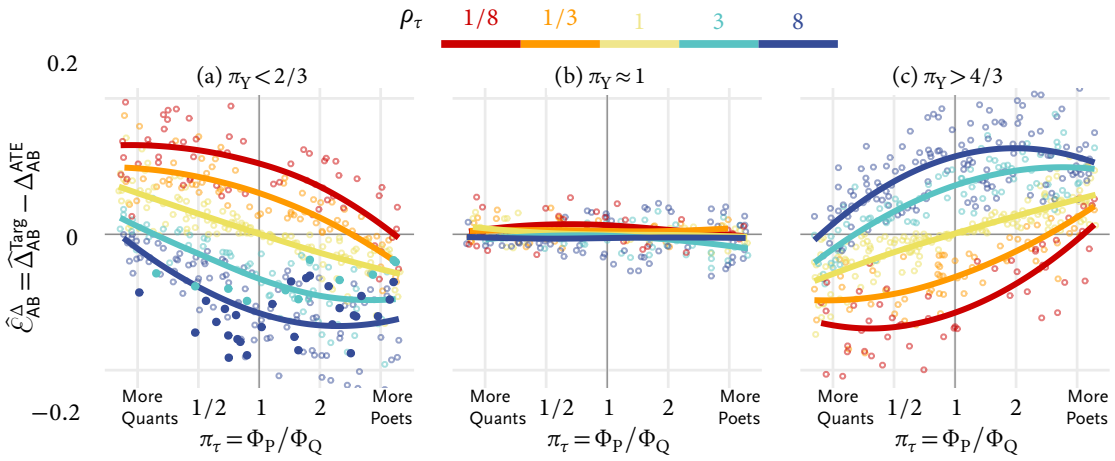
For ad B, the same response patterns and targeting policies have the opposite effect on bias ($\widehat{\mathcal{E}}^{\lambda}_{\mathrm{B}} < 0$). Unlike in the top panel where ρ_τ and π_τ both favored A, in the middle panel this same targeting policy of $\pi_\tau < 1$ (left side of x-axis) and $\rho_\tau = 1/8$ (red) creates two opposing effects on B. On the margin, targeting favors the weaker responding B-Quants (through $\pi_\tau < 1$), which in turn depresses $\hat{\lambda}^{\mathrm{Targ}}_{\mathrm{B}}$, but the divergent delivery aspect of the targeting policy favors targeting the better responding B-Poets (through $\rho_\tau < 1$), which increases $\hat{\lambda}^{\mathrm{Targ}}_{\mathrm{B}}$. As the mix becomes less dominated by Quants ($\pi_\tau < 1$ but increasing left to right), the lift estimate approaches the audience value ($\widehat{\mathcal{E}}^{\lambda}_{\mathrm{B}} = 0$). Then, as it becomes more dominated by Poets ($\pi_\tau \geq 1$), the targeted mix overestimates the true audience lift ($\widehat{\mathcal{E}}^{\lambda}_{\mathrm{B}} > 0$). As a result, the bias in the estimated difference in the lifts ($\widehat{\mathcal{E}}^{\Delta}_{\mathrm{AB}} = \widehat{\mathcal{E}}^{\lambda}_{\mathrm{A}} - \widehat{\mathcal{E}}^{\lambda}_{\mathrm{B}} = \widehat{\Delta}^{\mathrm{Targ}}_{\mathrm{AB}} - \Delta^{\mathrm{ATE}}_{\mathrm{AB}}$; bottom left of Fig. 7) will be even more extreme than the bias in each ad's estimated lift individually. This is because the lifts' biases are impacted by the same targeting strategy in opposite ways, with the same magnitude, which we show in the Appendix. *Subtracting* the two effects accentuates the bias under all targeting strategies (except, trivially, when $\pi_\tau = 1$). When users respond similarly on the margin ($\pi_{\mathrm{Y}} \approx 1$), divergent delivery ($\rho_\tau$) does not affect the bias $\widehat{\mathcal{E}}^{\Delta}_{\mathrm{AB}}$, which is why the colored lines overlap in the bottom row of Fig. 7.

Example 2 (Fig. 7, middle column; $\alpha_{\mathrm{Y}} > 1,\ \pi_{\mathrm{Y}} \approx 1,\ \rho_{\mathrm{Y}} = 1$) The middle column of Fig. 7 shows a case where users' responses are entirely homogeneous. While A is still stronger than B overall ($\alpha_{\mathrm{Y}} > 1$), now user types respond similarly not only in aggregate ($\pi_{\mathrm{Y}} \approx 1$), but also to each ad ($\rho_{\mathrm{Y}} = 1$). Without response heterogeneity from either π_{Y} or ρ_{Y}, all targeting decisions (all combinations of α_τ, π_τ and ρ_τ) equally cause no average bias in aggregate lifts ($\widehat{\mathcal{E}}^{\lambda}_{\mathrm{A}} \approx \widehat{\mathcal{E}}^{\lambda}_{\mathrm{B}} \approx 0$), and therefore, no bias in A/B difference in lifts ($\widehat{\Delta}^{\mathrm{Targ}}_{\mathrm{AB}} \approx \Delta^{\mathrm{ATE}}_{\mathrm{AB}}$, $\widehat{\mathcal{E}}^{\Delta}_{\mathrm{AB}} \approx 0$).

Example 3 (Fig. 8, right column; $\alpha_{\mathrm{Y}} > 1,\ \pi_{\mathrm{Y}} > 1,\ \rho_{\mathrm{Y}} = 8$) For a third example we turn to Fig. 8, whose panels all still have ad A is stronger than ad B on average ($\alpha_{\mathrm{Y}} > 1$). But now there is marginal user heterogeneity where Poets respond better than Quants to ads on average ($\pi_{\mathrm{Y}} > 1$). The top right panel describes the bias in lift ($\widehat{\mathcal{E}}^{\lambda}_{\mathrm{A}}$) for audiences with high user-ad response interaction ($\rho_{\mathrm{Y}} = 8$), where A-Poets and B-Quants have higher lift propensities than their "marginal" α_{Y} and π_{Y} would suggest. The different targeting policies (combinations of π_τ and ρ_τ) have a particularly large effect on the deviation between $\hat{\lambda}^{\mathrm{Targ}}_{\mathrm{A}}$ and $\lambda^{\mathrm{ATE}}_{\mathrm{A}}$. To see why, consider the most extreme targeting policy shown ($\pi_\tau = 4, \rho_\tau = 8$), where ad A is delivered more heavily to Poets, and those same A-Poets are exactly those who have the strongest response to ad A. Given that alignment where the best responding user-ad pair is also the most targeted (e.g., $\pi_{\mathrm{Y}} > 1$, $\rho_{\mathrm{Y}} > 1$, $\pi_\tau > 1$, and $\rho_\tau > 1$), the estimated lift of ad A for it's targeted mix will be higher than the ad's true average audience lift. Therefore, $\widehat{\mathcal{E}}^{\lambda}_{\mathrm{A}} > 0$.

Under the same targeting policy ($\pi_{\mathrm{Y}} = 4$, $\rho_\tau = 8$), ad B will be delivered to more Quants than Poets, and even more heavily to B-Quants. But the response rate of the B-Quants is affected by two opposing forces. The marginal effect π_{Y} points in one direction — Quants respond worse than Poets overall, and ad B is weaker than A, on average. But the user-ad response interaction effect in ρ_{Y} points in the other direction — the B-Quants have a greater response rate than the marginal effects alone dictate. The B-Poets' response rate exhibits similar offsetting

Figure 9: Simulated $\hat{\mathcal{E}}^{\Delta}_{AB}$ for $\rho_Y = 1,\ \alpha_Y > 1$



Note to Figure 9: Solid dots in Fig. 9a indicate audiences that meet the criteria for a Simpson's reversal.

forces: Poets respond better, but *B*-Poets have an even lower response to *B*. As a result, all targeting decisions create minimal bias for ad *B* ($\hat{\mathcal{E}}^{\lambda}_{B}$ is small). Therefore, dominated by the $\hat{\mathcal{E}}^{\lambda}_{A}$, the bias in estimated difference in lifts, $\hat{\mathcal{E}}^{\Delta}_{AB}$, will still be biased for all unbalanced targeting policies ($\pi_Y \neq 1, \rho_\tau \neq 1$).

In Fig. 9, three panels illustrate the interesting case of audiences with no response heterogeneity ($\rho_Y = 1$) even though there may be differences across user types in marginal responses (π_Y). As long as the ads differ in average effects ($\alpha_Y > 1$), targeting can affect bias in the presense of marginal user response (π_Y) alone. But when there is no heterogeneity at all ($\rho_Y = 1$, $\pi_Y \approx 1$; Fig. 9b), targeting cannot generate a bias in A/B difference in lifts ($\widehat{\Delta}^{Targ}_{AB} \approx \Delta^{ATE}_{AB}$, $\hat{\mathcal{E}}^{\Delta}_{AB} \approx 0$).

Example 4 (Fig. 9c; $\alpha_Y > 1,\ \pi_Y > 1,\ \rho_Y = 1$) For audiences where Poets respond better than Quants overall, we consider the bias when targeting skews the mix of targeted users towards more Poets than Quants ($\pi_\tau > 1$; right side of *x*-axis) and engages in a divergent delivery policy that results in additional Poets seeing *A* and Quants seeing *B* ($\rho_\tau = 8$; blue). Aggregate lift of *A* is overestimated as the most responsive *A*-Poets are "doubly favored" by a targeting policy with high π_τ and ρ_τ. At the same time, the lift of *B* is only slightly underestimated due to the offsetting forces (a high π_τ favors *B*-Poets, but a high ρ_τ favors *B*-Quants). Thus, the bias in the A/B difference is positive, but tapers out and declines as targeting of Poets becomes more extreme.

Example 5 (Fig. 9a; $\alpha_Y > 1,\ \pi_Y < 1,\ \rho_Y = 1$; Simpson's reversal) The *filled* circles in Fig. 9a identify audiences whose simulated estimates meet the conditions for a Simpson's reversal (Sec. 4.1.1), where aggregation bias can cause $\widehat{\Delta}^{Targ}_{AB}$ to have a different sign from the Δ^{ATE}_{AB} the advertiser is trying to infer. Compared to Fig. 9c ($\pi_Y > 4/3$), Fig. 9a ($\pi_Y > 4/3$) has patterns of bias that are rotated around the origin 180°. Because $\pi_Y < 2/3$ and $\rho_Y = 1$, it follows that Quants respond to ad *B* much better that Poets (Factor Eq. 29.2 is positive and large). And because

$\alpha_Y > 1$ and $\rho_Y = 1$, A is the stronger ad among both user types (Factors Eq. 29.1a and Eq. 29.1b are both positive, but not necessarily large). A condition for Simpson's reversal to occur is "overtargeting." When $\pi_\tau > 1$ and $\rho_\tau > 1$ (the right end of the x-axis on the blue line in Fig. 9a), the algorithm is engaging in a divergent delivery targeting policy that exposes more of the worse responding Poets to the stronger ad A, lowering the estimate of the aggregate lift of A, relative to its true value in the audience. At the same time, the mix of users targeted with ad B contains more of the better responding Quants than are in the audience, so the effect of ad B is overestimated. If this divergence in mixtures is strong enough, then σ_{QA} will be small relative to σ_{QB}, making Eq. 29.3 (and the entire LHS of Eq. 29) large. So as long as the difference in the true ads effects, within user types, is small, the RHS of Eq. 29 will be small enough that the estimated aggregated lift for B will be higher than A, even though both user types respond better to A than B.

General insights from the simulation Looking across the bottom rows of Figs. 7 and 8, and Fig. 9, we summarize several general relationships that generate distinct patterns of bias in the A/B difference ($\hat{\mathcal{E}}^{\Delta}_{AB}$).

- Variation in π_τ along the x-axes reflects deviation in the mix from σ_{XZ} from γ_X. Vertical distances of the colored lines from the yellow line ($\rho_\tau = 1$) reflects separation between σ_{PA} and σ_{PB}.
- In the absence of both types of user-level differences ($\pi_Y = 1$ and $\rho_Y = 1$), no amount of targeting will create bias $\hat{\mathcal{E}}^{\Delta}_{AB}$. Targeting policies generate bias in experimental results because of heterogeneity, as either π_Y or ρ_Y deviates from 1.
- Even when the audience exhibits user-ad response interaction, the absence of marginal user response heterogeneity ($\pi_Y = 1$) eliminates the impact of divergent delivery ($\rho_\tau \neq 1$) on bias, but does not entirely eliminate the bias caused by marginal user targeting.
- Even when the targeting policy includes divergent delivery, the absence of marginal targeting by user ($\pi_\tau = 1$) eliminates the impact of user-ad response interactions on bias ($\rho_Y \neq 1$), but does not entirely eliminate the bias caused by marginal user heterogeneity.
- More extreme values of bias ($\hat{\mathcal{E}}^{\Delta}_{AB}$) appear when there is alignment across parameters; e.g., when user response is strongest among Poets ($\pi_Y > 1$), for ad A ($\alpha_Y > 1$), and especially among A-Poets and B-Quants ($\rho_Y > 1$), and when targeting favors the Poets ($\pi_\tau > 1$), ad A ($\alpha_\tau > 1$), and especially A-Poets and B-Quants ($\rho_\tau > 1$).

5 Disabling divergent delivery

The simulation shows that the bias is largely determined by the particular relationship between user heterogeneity across types (π_Y) and divergent delivery (ρ_τ). This means that in the presence of response heterogeneity — across users with or without an interaction with the ad — there is nothing an *advertiser* can do by itself to mitigate the bias in A/B difference estimates caused by algorithmic targeting. However, the platform could help by applying

different targeting policies during an experiment than it does for a rollout of the ad campaign. The most extreme remedy to the bias would be to stop targeting altogether, exposing all users to their assigned ads with a common probability $\Phi_{XZ} = \widetilde{\Phi}$ (i.e., setting $\pi_\tau = 1$ and $\rho_\tau = 1$). A more realistic and moderate alternative would be to target all ads in the experiment collectively, as a single unit, to a single mix of users. Such a targeting policy, we say, is *disabling divergent delivery*. This involves setting $\rho_\tau = 1$, so $\Phi_{XA} = \Phi_{XB} = \Phi_X$, and $\sigma_{XA} = \sigma_{XB} = \sigma_X$, but still allowing the mix of users assigned to *any* of the experimental treatments to be different than the mix of the audience ($\sigma_X \neq \gamma_X$). This design permutes the conceptual "order of operations" from the tree in Fig. 2 by *first* targeting users to the campaign, and *then* randomly assigning users to an ad treatment. The later randomization of users to arms in the A/B/n test with holdout remains unchanged.

Fig. 10 shows the *incremental bias* generated by divergent delivery ($\hat{\mathcal{E}}^{\Delta}_{AB} \mid \pi_\tau, \rho_\tau$), relative to what the bias would have been had divergent delivery been disabled ($\hat{\mathcal{E}}^{\Delta}_{AB} \mid \pi_\tau, \rho_\tau = 1$).[17] In this figure, we fix audiences to have high user-ad response interaction ($\rho_Y = 8$), so we can examine the large effect of π_τ and ρ_τ, moderated by π_Y. For a given level of divergent delivery (ρ_τ), the magnitude of the incremental bias is greatest when there is high heterogeneity in responsiveness between Poets and Quants (π_Y; extreme ends of the color scale), yet each type is nevertheless equally likely to be targeted overall ($\pi_\tau = 1$ on the x-axis). For example, when $\rho_\tau = 8$ (right panel), divergent delivery favors the *A*-Poets, who are already highly responsive ($\rho_Y = 8$ in all panels). So when user heterogeneity shows Poets respond stronger than Quants overall as well, (e.g., $\pi_Y = 4$, red), divergent delivery creates more of an overestimation of ad *A*'s lift relative to *B*'s lift than the overestimation that would happen if the algorithm similarly targeted Poets more than Quants overall ($\pi_\tau > 1$), but did not use divergent delivery ($\rho_\tau = 1$).[18]

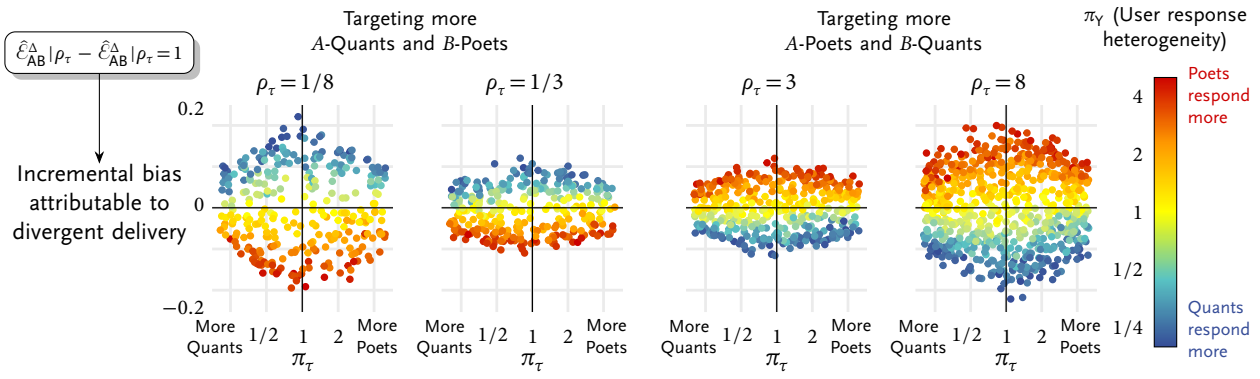
While disabling divergent delivery may reduce bias in causal inference, it comes at a cost. Fig. 11 quantifies the economic impact of divergent delivery in terms of the percentage difference in conversions under targeting policies with and without divergent delivery *for the entire audience*. The incremental conversions correspond to value for the platform (as advertisers may charge advertisers per conversion) and for the advertiser (as conversions are revenue-relevant events).

As expected, the incremental conversions from a targeting policy that employs divergent delivery, compared to one without, is greater when targeting "points in the same direction" as the audience's user-ad response interaction (i.e., $\rho_\tau < 1, \rho_Y < 1$; or $\rho_\tau > 1, \rho_Y > 1$). But that turns into a loss when the algorithm is "mistargeting" ($\rho_\tau < 1, \rho_Y > 1$; or $\rho_\tau > 1, \rho_Y < 1$). Matching the platform's divergent delivery policy to the targeted users' ad response heterogeneity will bring in more money, but also incrementally increase the bias in the estimate of the

[17] Incremental bias is a "quadruple-diff" value. Each panel's ρ_τ value in Fig. 10 matches the ρ_τ in the color scale of the bias plots in the bottom row of Figs. 7 and 8, and in Fig. 9. Thus, the incremental bias in Fig. 10 is equivalent to the vertical distance between each colored line (ρ_τ) and the yellow line ($\rho_\tau = 1$) in Figs. 7 to 9.

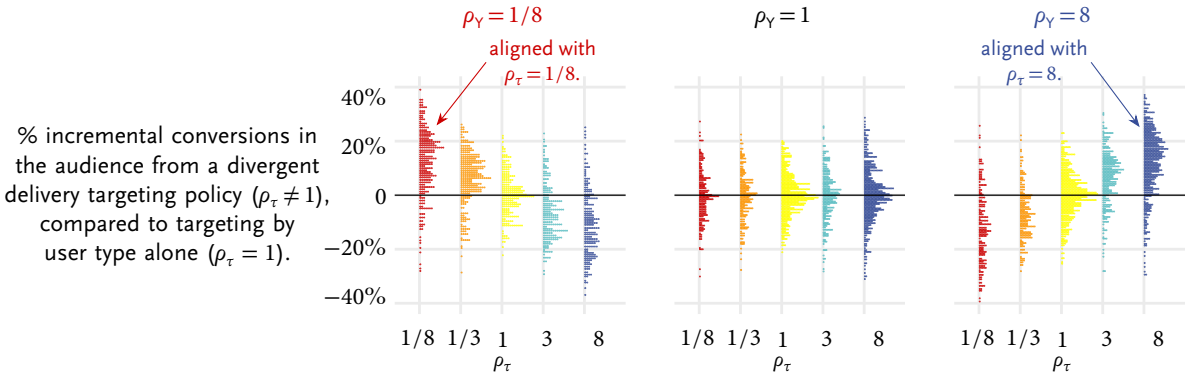
[18] At the extremes of the x-axis, the algorithm is targeting predominantly one user type or the other, so divergent delivery doesn't affect the bias much. If $\pi_\tau \to \infty$, then $\sigma_{PA} \to 1$ and $\sigma_{PB} \to 1$. Therefore, $\lambda^{Targ}_A \to \lambda_{PA}$ and $\lambda^{Targ}_B \to \lambda_{PB}$. So while estimates $\hat{\lambda}^{Targ}_A$, $\hat{\lambda}^{Targ}_B$, and $\widehat{\Delta}^{Targ}_{AB}$ are affected by extreme values of π_τ, they would not depend on the divergent delivery (ρ_τ). The reduction in bias by disabling divergent delivery would be minimal.

Figure 10: Incremental Bias Attributable to Divergent Delivery



Note to Figure 10: The y-axis is the difference in divergent delivery bias ($\hat{\mathcal{E}}^{\Delta}_{AB}$) between two targeting policies: one with divergent delivery ρ_τ (noted at the top of each panel), and one without ($\rho_\tau = 1$). All panels show audiences with $\rho_Y = 8$. The effects are persistent, but attenuated, for smaller values of $\rho_Y > 1$, and they reverse for reciprocal values $\rho_Y < 1$.

Figure 11: Incremental Conversions in an Audience Generated by Divergent Delivery



A/B difference. Efforts to reduce bias by jointly targeting the entire set of ads in the campaign to a common set of users ($\rho_\tau = 1$) would not be as profitable as targeting policies that employ divergent delivery. The economic value from divergent delivery can explain why we find ourselves in an equilibrium where platforms do not offer an option to disable divergent delivery, and advertisers accept this. Even during an experiment, advertisers still pay for the ads that are shown, so both they and the platform are seeking a return on that expenditure. Also, one of the reasons advertisers provide experimental tools is to show off how well the platform generates value. The targeting algorithm is part of that, but as we explain in 3.3.1, the relative contribution of the targeting algorithm cannot be identified. So even if the advertiser wants to reduce experimental bias from targeting, it may not always be in the publisher's interest to let them.

6 Discussion

While the idea of divergent delivery and its consequences for conducting causal inference on ad creatives with online experiments are not entirely new to marketing research community, this paper is the first to formally define

and analyze it mathematically. Beyond establishing the framework through which future researchers can study the effects of algorithmic targeting, our contribution to the literature revolves around our explanation of how algorithmic ad targeting generates bias: how the mixes of *targeted* users differ from the audience, and across ads, under particular patterns of heterogeneity. Using available online experimentation tools, an advertiser cannot take the difference in lift between ad *A* and ad *B* and make a causal interpretation about preferences among the the audience. The reason is that divergent delivery of ads to a heterogeneous audience prevents advertisers from separately distinguishing the effects of their ad creatives on users from the effects of how the algorithm selects users to see each ad. The confound arises because platforms report results that are aggregated across the same unobserved user types that the platform's algorithm utilizes for targeting. So, For *between-ad comparisons*, this problem is not resolved by random holdout approaches that are otherwise effective for testing the effectiveness of a single ad.

We have some simple advice for advertisers and researchers who care about inferences on a predefined population, but are considering running so-called "randomized" experiments on platforms that target ads to users: *Don't*. Or at least, be wary. Online publishers add value for advertisers by delivering different ads to different types of users. Experimentation tools are one way for publishers to demonstrate that value, and to help advertisers optimize that value. The same targeting algorithm that improves lift for a single ad simultaneously distorts estimates of the difference in lift between ads. However, our advice for advertisers whose concerns are to predict which version of an ad will "do best" in a non-experimental campaign is different: *Carry on*. Still, even those advertisers should be aware that A/B comparisons are measuring more than the effects from creative elements, including a combination of ad effects, targeting effects, and their interaction, while the advertisers cannot detect component's effect separately.

Perhaps the result with the biggest managerial implication is the potential for a Simpson's reversal. Targeting bias is not just a question of magnitude of effects, but also the *sign*. Algorithms that *overtarget* ads to users when the cross-ad effects are actually quite small are especially prone to lead experimenters into the Simpson's reversal trap. Advertisers of all kinds — commercial, academic, and governmental — need to be aware of this possibility, and how an unobservable Simpson's reversal can manifest when heterogeneity, targeting policies, and aggregation of results are all aligned in a certain way. Although our analysis uses the simple two-type, two ad case, concerns about a Simpson's reversal are still quite relevant in practice, when the number of user types is much larger. In the general case with n_{X} user types, a Simpson's reversal happens when, $\lambda_{x_1\mathrm{A}} > \lambda_{x_1\mathrm{B}}$, $\lambda_{x_2\mathrm{A}} \geq \lambda_{x_2\mathrm{B}}$, … , $\lambda_{x_{n_{\mathrm{X}}}\mathrm{A}} \geq \lambda_{x_{n_{\mathrm{X}}}\mathrm{B}}$, (with strict inequality for at least one type), but $\hat{\lambda}_{\mathrm{A}}^{\mathrm{Targ}} < \hat{\lambda}_{\mathrm{B}}^{\mathrm{Targ}}$ in aggregate. While a "pure" Simpsons reversal cannot arise if any of the $\lambda_{x_j\mathrm{B}} > \lambda_{x_j\mathrm{A}}$ for at least one user type, the basic reversal can still occur among many other subsets of users with the same preferred ad. The fact that there is any bias at all places the advertiser at risk of making decisions based on data from a subset of users that is not representative of the population of interest.

Also, this paper is stands out in that it studies bias in online experiments from the point of view of the advertiser. The research streams most relevant to our setting have largely been conducted with the involvement of, and from the point of view of, the publishers themselves: Google in the case of Johnson et al. (2017a) and Johnson et al. (2017b); Facebook for Gordon et al. (2019); Yahoo! for Lewis and Reiley (2014); and Linkedin for Xu et al. (2015). Even platform-sponsored work that is independent of a particular platform work (Bakshy et al. 2014) nevertheless takes a platform-centric view, concerned with platform's experimental design issues. But if platforms are going to encourage advertisers to conduct experiments using their experimental tools, we need to consider that the advertiser may not be learning what it thinks it is learning. This is certainly true in the cases of academic researchers acting as advertisers, and in many commercial market research contexts as well.

If they were so inclined, publishers *could* offer advertisers the option of an experimental design that would not target users based on specific ads while an experiment is ongoing (as in Sec. 5). But disabling divergent delivery comes at a cost, and does not show the algorithm at work in the best possible light. Or publishers *could* reduce aggregation bias by providing advertisers more finely-grained reports of results, essentially converting some the unobservable dimensions of heterogeneity to observable (which is a good path for future study). But disaggregating data has serious privacy implications. So we ask: why would the publisher even want to reduce the bias in the A/B difference? This bias is defined as the difference between the A/B comparison among users targeted by the platform, and the A/B comparison among a broader population the advertiser is studying. The publisher wants the *targeted* A/B difference to be accurate because it makes more money when the advertiser runs good ads *on that publisher's platform*. But inferences about the effectiveness of creative elements among a broader population can be generalizable to other media channels. For example, the advertiser can use information gleaned from tests of different versions of ad copy to develop creative material to be run on competitors' platforms, or even for offline advertising. One could consider additional services that are aligned with both parties' interests. But additional research into publishers' incentives to keep experimental results applicable and relevant only for their platform, and not useful for off-platform deployment, is warranted.

This paper is not about an external validity issue. As a field experiment, these ad tests carry stronger external validity than a "lab experiment" in consumer psychology research, but the defense of external validity is similar. There is no direct evidence or claim that results generalize outside of the platform, the experiment, or that moment in time. But more so than a lab study, an online ad experiment in a platform with an ad market enables the advertiser to better define its reference population, selecting criteria for including potential subjects to be eligible for targeting and eventual exposure. Results are generalizable only to that reference population.

As with any quantitative research there are tradeoffs between model simplicity and the ability to generate meaningful insights. We were careful that our results would not depend on stylized examples or heavily parametrized models. All quantities in the model are averages, differences or ratios, and all random samples in the simulation

come from uniform or log-uniform distributions. We also recognize that the orders of magnitude of targeting and response probabilities are larger than one might see in practice, but the same patterns would arise just by using larger simulated audiences. We bounded the scope of this paper to the case of two user types because we think it is the best way to present our framework and results in words, numbers, and pictures on a two-dimensional page.

For future work, our framework opens the door for others to consider directly addressing complex experimental designs (factorial) and analyses (regression). Other opportunities for future research include closer attention to availability bias, which we assumed away by setting $V = 1$ in the simulation. We would also be interested in how well our results and insights hold up in the cases of large n_X and n_Z.

References

Ali, M., P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke (2019). "Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes." *Proceedings of the ACM on Human-Computer Interaction*, 3(199):1–30.

Baker, S. G. and B. S. Kramer (2001). "Good for Women, Good for Men, Bad for People: Simpson's Paradox and the Importance of Sex-Specific Analysis in Observational Studies." *Journal of Women's Health and Gender-Based Medicine*, 10(9):867–872.

Bakshy, E., D. Eckles, and M. S. Bernstein (2014). "Designing and Deploying Online Field Experiments." *WWW '14 Proceedings of the 23rd International Conference on World Wide Web*. 283–292.

Blyth, C. R. (1972). "On Simpson's Paradox and the Sure-Thing Principle." *Journal of the American Statistical Association*, 67(338):364–366.

Cecere, G., C. Jean, M. Manant, and C. Tucker (2018). "Computer Algorithms Prefer Headless Women." *2018 MIT CODE : Conference on Digital Experimentation*. URL: https://hal.archives-ouvertes.fr/hal-02333913.

Eckles, D., B. R. Gordon, and G. A. Johnson (2018). "Field Studies of Psychologically Targeted Ads Face Threats to Internal Validity." *Proceedings of the National Academy of Sciences*, 115(23):E5254–E5255.

Feit, E. M. and R. Berman (2019). "Test & Roll: Profit-Maximizing A/B Tests." *Marketing Science*, 38(6):1038–1058.

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., 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.

Hao, K. (2021). "Facebook's Ad Algorithms are Still Excluding Women from Seeing Jobs." *MIT Technology Review*. URL: https://www.technologyreview.com/2021/04/09/1022217/facebook-ad-algorithm-sex-discrimination/.

Johnson, G. A. (2020). "Inferno: A Guide to Field Experiments in Online Display Advertising". Working paper. Boston University. SSRN:3581396.

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

Johnson, G. A., R. A. Lewis, and E. I. Nubbemeyer (2017b). "The Online Display Ad Effectiveness Funnel and Carryover: Lessons from 432 Field Experiments". Working paper. SSRN:2701578.

Kupor, D. and K. Laurin (2020). "Probable Cause: The Influence of Prior Probabilities on Forecasts and Perceptions of Magnitude." *Journal of Consumer Research*, 46(5):833–852.

Kupor, D., K. Laurin, and J. Levav (2015). "Anticipating Divine Protection? Reminders of God Can Increase Nonmoral Risk Taking." *Psychological Science*, 26(4):374–384.

Lewis, R. A., J. M. Rao, and D. H. Reiley (2011). "Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising." *WWW '11 Proceedings of the 20th International Conference on World Wide Web*.

Lewis, R. A. and D. H. Reiley (2014). "Online Ads and Offline Sales: Measuring the Effects of Retail Advertising via a Controlled Experiment on Yahoo!" *Quantitative Marketing and Economics*, 12:235–266.

Lodish, L. M., M. Abraham, S. Kalmenson, J. Livelsberger, B. Lubetkin, B. Richardson, and M. E. Stevens (1995). "How TV Advertising Works: A Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments." *Journal of Marketing Research*, 32(2):125–139.

Matz, S. C., M. Kosinski, G. Nave, and D. J. Stillwell (2017). "Psychological Targeting as an Effective Approach to Digital Mass Persuasion." *Proceedings of the National Academy of Sciences*, 114(48):12714–12719.

Orazi, D. C. and A. C. Johnston (2020). "Running Field Experiments Using Facebook Split Test." *Journal of Business Research*, 118:189–198.

Pearl, J. (2014). "Comment: Understanding Simpson's Paradox." *The American Statistician*, 68(1):8–13.

Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 66(5):688–701.

Shapiro, B., G. J. Hitsch, and A. Tuchman (2020). "Generalizable and Robust TV Advertising Effects". Working paper. National Bureau of Economic Research. NBER:27684.

Simpson, E. H. (1951). "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society B*, 13(2):238–241.

Xu, Y., N. Chen, A. Fernandez, O. Sinno, and A. Bhasin (2015). "From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks." *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2227–2236.

Appendix

Relationship between ρ_τ and σ_{XZ} in Fig. 4 This section provides mathematical support for the effects of targeting policies on the mixes of targeted users, as illustrated in Fig. 4. Rearranging Eq. 12, Eq. 13, and Eq. 15, respectively, $\Phi_{XZ} = \frac{\sigma_{XZ}\Phi_Z}{\gamma_X}$, $\Phi_A = \alpha_\tau \Phi_B$, and $\sigma_{PB} = \frac{\sigma_{PA}}{\rho_\tau + (1-\rho_\tau)\,\sigma_{PA}}$. Substituting these terms into Eq. 14 and rearranging gives us an expression for the posterior odds a targeted A user is a Poet.

$$\frac{\sigma_{PA}}{1-\sigma_{PA}} = \pi_\tau \frac{\gamma_P}{1-\gamma_P} G, \quad \text{where} \quad G = \left[\frac{\alpha_\tau \zeta_A \left(\sigma_{PA} + \rho_\tau \left(1-\sigma_{PA}\right)\right) + \left(1-\zeta_A\right)\rho_\tau}{\alpha_\tau \zeta_A \left(\sigma_{PA} + \rho_\tau \left(1-\sigma_{PA}\right)\right) + \left(1-\zeta_A\right)} \right] \tag{A.1}$$

The numerator and denominator of G differ only in the final terms of each. If $\rho_\tau = 1$, then $G = 1$, so in the absence of divergent delivery, the posterior odds $\frac{\sigma_{PA}}{1-\sigma_{PA}}$ that a targeted A-user is a Poet is linear in π_τ. Also when $\rho_\tau = 1$, $\sigma_{PA} = \sigma_{PB}$, so changing the overall mix of Poets and Quants (π_τ) affects the targeted mixes of both ads in the same proportions. This is represented in the top two rows of Fig. 4 where a change in π_τ vertically shifts the two ads'

targeting ovals *in tandem*. In the special case of $\rho_\tau = 1$ and $\pi_\tau = 1$ (Fig. 4, top two rows, center column), there is no targeting by user type at all, so the proportions of blue inside the ovals ($\sigma_{PA} = \sigma_{PB}$) are the same as in the audience (γ_P). Divergent delivery ($\rho_\tau > 1$, then $G > 1$) provides an additional "bump" to $\frac{\sigma_{PA}}{1-\sigma_{PA}}$ by a factor of G by targeting additional A-Poets. Because $\frac{\sigma_{PB}}{1-\sigma_{PB}} = \frac{1}{\rho_\tau}\frac{\sigma_{PA}}{1-\sigma_{PA}}$ (Eq. 15), the proportion of Poets among users targeted with A would be higher than those targeted with B ($\sigma_{PA} > \sigma_{PB}$). We see this effect of divergent delivery in the bottom two rows of Fig. 4 ($\rho_\tau = 8$) where the ovals separate vertically, and the blue proportions of the A and B ovals diverge.

Effects of changing mix on bias To better understand the effect of divergent delivery on the bias, we consider the effects from perturbations of σ_{PA} and σ_{PB}. Differentiating Eq. 25,

$$\frac{\partial \mathcal{E}^{\Delta}_{AB}}{\partial \sigma_{PA}} = \lambda_{PA} - \lambda_{QA} \qquad \frac{\partial \mathcal{E}^{\Delta}_{AB}}{\partial \sigma_{PB}} = \lambda_{QB} - \lambda_{PB} \tag{A.2}$$

where $\lambda_{PA} - \lambda_{QA}$ and $\lambda_{PB} - \lambda_{QB}$ are related through π_Y and α_Y. For simplicity, let's assume for all X and Z that $\Theta^{(0)}_{XZ} = 0$, so $\lambda_{XZ} = \Theta^{(1)}_{XZ}$. Solving Eqs. 26 and 27 for λ_{PA} and λ_{QA},

$$\lambda_{PA} = \lambda_{PB}\left[\frac{\gamma_P\left(1+\zeta_A\left(\alpha_Y\pi_Y - 1\right)\right) + \zeta_A - 1}{\zeta_A\left(1+\gamma_P\left(\pi_Y - 1\right)\right)}\right] + \lambda_{QB}\left[\frac{\pi_Y\left(\left(1-\gamma_P\right)\left(1+\zeta_A\left(\alpha_Y - 1\right)\right)\right)}{\zeta_A\left(1+\gamma_P\left(\pi_Y - 1\right)\right)}\right] \tag{A.3}$$

$$\lambda_{QA} = \lambda_{PB}\left[\frac{\gamma_P\left(1+\zeta_A\left(\alpha_Y - 1\right)\right)}{\zeta_A\left(1+\gamma_P\left(\pi_Y - 1\right)\right)}\right] + \lambda_{QB}\left[\frac{\pi_Y\left(\gamma_P\left(\alpha_Y - 1\right)\right) + \alpha_Y\left(\zeta_A\left(1-\gamma_P\right)\right)}{\zeta_A\left(1+\gamma_P\left(\pi_Y - 1\right)\right)}\right] \tag{A.4}$$

Substituting into Eq. A.2,

$$\frac{\partial \mathcal{E}^{\Delta}_{AB}}{\partial \sigma_{PA}} = \lambda_{PB}\left[\frac{\zeta_A\alpha_Y\gamma_P\left(\pi_Y - 1\right) + \zeta_A - 1}{\zeta_A\left(1+\gamma_P\left(\pi_Y - 1\right)\right)}\right] + \lambda_{QB}\left[\frac{\zeta_A\alpha_Y\left(1-\gamma_P\right)\left(\pi_Y - 1\right) + \left(1-\zeta_A\right)\pi_Y}{\zeta_A\left(1+\gamma_P\left(\pi_Y - 1\right)\right)}\right] \tag{A.5}$$

$$\frac{\partial \mathcal{E}^{\Delta}_{AB}}{\partial \sigma_{PB}} = \lambda_{QB} - \lambda_{PB} \tag{A.6}$$

In the case when aggregate response rates for Poets and Quants are equal ($\pi_Y = 1$), Eqs. A.5 and A.6 reduce to

$$\frac{\partial \mathcal{E}^{\Delta}_{AB}}{\partial \sigma_{PA}} = \left(\frac{1-\zeta_A}{\zeta_A}\right)\cdot\left(\lambda_{QB} - \lambda_{PB}\right) = \left(\frac{1-\zeta_A}{\zeta_A}\right)\cdot\frac{\partial \mathcal{E}^{\Delta}_{AB}}{\partial \sigma_{PB}} \tag{A.7}$$

Equation A.7 shows that in this case, incrementally targeting more Poets, but only those assigned to A, will move the bias in the same direction as if targeting more Poets with ad B. The direction of bias depends on which user type has the greater lift for ad B. Under a divergent delivery policy, when the algorithm targets a higher proportion of Poets among the A users, it is also more likely to target a *lower* fraction Poets among the B users. If the initial assignment of users to ads is balanced ($\zeta_A = 1/2$), the magnitudes of the opposing forces are the same as well. These offsetting effects explain the bottom row of Fig. 7, where changing ρ_τ does not affect the bias (no vertical shift) when user types are homogeneous in their response to all ads in the campaign ($\pi_Y = 1$).. However, if initial randomization of the audience to ads were not balanced, divergent delivery might still generate bias in the estimated $\widehat{\Delta}_{AB}$ (Eq. A.7).

Web Appendix

The simulation has four levels:

- the parameter ratios from which the audience profile is generated (α_τ, π_τ, ρ_τ, α_Y, π_Y, and ρ_Y);
- the audience profile itself ($\Theta^{(0)}_{XZ}$, $\Theta^{(1)}_{XZ}$, and Φ_{XZ} for each $X \in \{P, Q\}$ and $Z \in \{A, B, C\}$);
- the user-level potential outcomes ($Y^{(1)}_Z$ and $Y^{(0)}_Z$); and
- actions by the platform (targeting decisions τ, and test arm assignments R).

The following algorithm generates audience parameters and profiles for a given $(\rho_\tau,\ \rho_Y)$ pair.

- Set bounds and initial values.
  1. Set lower bounds $\underline{\alpha}_\tau = 1/2$, $\underline{\alpha}_Y = 1/4$, $\underline{\pi}_\tau = 1/4$, and $\underline{\pi}_Y = 1/4$; and upper bounds $\overline{\alpha}_\tau = 2$, $\overline{\alpha}_Y = 4$, $\overline{\pi}_\tau = 4$, and $\overline{\pi}_Y = 4$.
  2. For $X \in \{P, Q\}$, set lower and upper bounds $\underline{\Phi}_{XC} = \underline{\Theta}^{(1)}_{XC} = .02$ and $\overline{\Phi}_{XC} = \overline{\Theta}^{(1)}_{XC} = .04$.
  3. Set $\widetilde{\Phi} = .2$ and $\widetilde{\Theta}^{(1)} = .2$.
- Sample and set the following elements of the audience profile.
  4. Sample $\Phi_{PC} \sim \text{Unif}\left(\underline{\Phi}_{PC}, \overline{\Phi}_{PC}\right)$, $\Theta^{(1)}_{PC} \sim \text{Unif}\left(\underline{\Theta}^{(1)}_{PC}, \overline{\Theta}^{(1)}_{PC}\right)$, $\Phi_{QC} \sim \text{Unif}\left(\underline{\Phi}_{QC}, \overline{\Phi}_{QC}\right)$, and $\Theta^{(1)}_{QC} \sim \text{Unif}\left(\underline{\Theta}^{(1)}_{QC}, \overline{\Theta}^{(1)}_{QC}\right)$.
  5. Set $\Theta^{(0)}_{PA} \leftarrow \Theta^{(1)}_{PC}$, $\Theta^{(0)}_{QA} \leftarrow \Theta^{(1)}_{QC}$, $\Theta^{(0)}_{PB} \leftarrow \Theta^{(1)}_{PC}$, $\Theta^{(0)}_{QB} \leftarrow \Theta^{(1)}_{QC}$, $\Theta^{(0)}_{PC} \leftarrow \Theta^{(1)}_{PC}$, and $\Theta^{(0)}_{QC} \leftarrow \Theta^{(1)}_{QC}$.
- Sample marginal ratios α_τ, α_Y, π_τ, and π_Y.[W1]
  6. Sample $\alpha_\tau \sim \log_2\text{Unif}\left(\underline{\alpha}_\tau, \overline{\alpha}_\tau\right)$ and $\alpha_Y \sim \log_2\text{Unif}\left(\underline{\alpha}_Y, \overline{\alpha}_Y\right)$.
  7. If $\Phi_{PC} + \Phi_{QC} < 6\widetilde{\Phi} - 1$, then adjust $\underline{\pi}_\tau \leftarrow \max\left(\underline{\pi}_\tau,\ 6\widetilde{\Phi} - \Phi_{PC} - \Phi_{QC} - 1\right)$ and $\overline{\pi}_\tau \leftarrow \min\left(\overline{\pi}_\tau,\ \frac{1}{6\widetilde{\Phi} - \Phi_{PC} - \Phi_{QC} - 1}\right)$.
  8. If $\Theta^{(1)}_{PC} + \Theta^{(1)}_{QC} < 6\widetilde{\Theta}^{(1)} - 1$, then adjust $\underline{\pi}_Y \leftarrow \max\left(\underline{\pi}_Y,\ 6\widetilde{\Theta}^{(1)} - \Theta^{(1)}_{PC} - \Theta^{(1)}_{QC} - 1\right)$ and $\overline{\pi}_Y \leftarrow \min\left(\overline{\pi}_Y,\ \frac{1}{6\widetilde{\Theta}^{(1)} - \Theta^{(1)}_{PC} - \Theta^{(1)}_{QC} - 1}\right)$.
  9. Sample $\pi_\tau \sim \log_2\text{Unif}\left(\underline{\pi}_\tau, \overline{\pi}_\tau\right)$ and $\pi_Y \sim \log_2\text{Unif}\left(\underline{\pi}_Y, \overline{\pi}_Y\right)$
- Solve for the remaining elements of the audience profile.[W2]

  10. Set the following intermediate values.

$$S_\tau \leftarrow \sqrt{\left(\alpha_\tau\pi_\tau - 1\right)^2 + \left(\alpha_\tau - \pi_\tau\right)^2\rho_\tau^2 + 2\rho_\tau\left(\alpha_\tau\pi_\tau\left(\alpha_\tau + \pi_\tau + 4\right) + \alpha_\tau + \pi_\tau\right)}$$

$$S_Y \leftarrow \sqrt{\left(\alpha_Y\pi_Y - 1\right)^2 + \left(\alpha_Y - \pi_Y\right)^2\rho_Y^2 + 2\rho_Y\left(\alpha_Y\pi_Y\left(\alpha_Y + \pi_Y + 4\right) + \alpha_Y + \pi_Y\right)}$$

[W1] To sample a random variable $y \sim \log_2\text{Unif}\left(a, b\right)$, first sample $y^* \sim \text{Unif}\left(\log_2 a, \log_2 b\right)$, and set $y = 2^{y^*}$.

[W2] In Steps 11 and 12, dividing by F_τ and F_Y from Step 10 creates removable discontinuities at $\rho_\tau = 1$ and $\rho_Y = 1$. Adding a small value like 10^{-10} to ρ_τ and ρ_Y is a sufficient remedy.

$$F_\tau \leftarrow (\alpha_\tau + 1)\,(\pi_\tau + 1)\,(\rho_\tau - 1)$$
$$F_Y \leftarrow (\alpha_\mathrm{Y} + 1)\,(\pi_\mathrm{Y} + 1)\,(\rho_\mathrm{Y} - 1)$$

11. Set the remaining targeting probabilities.

$$\Phi_{\mathrm{PA}} \leftarrow \frac{2\widetilde{\Phi}}{F_\tau}\left(\rho_\tau(\alpha_\tau + \pi_\tau + 2\alpha_\tau\pi_\tau) - \alpha_\tau\pi_\tau - S_\tau + 1\right)$$
$$\Phi_{\mathrm{PB}} \leftarrow \frac{2\widetilde{\Phi}}{F_\tau}\left(\pi_\tau(\rho_\tau - 2) - \alpha_\tau(\pi_\tau + \rho_\tau) + S_\tau - 1\right)$$
$$\Phi_{\mathrm{QA}} \leftarrow \frac{2\widetilde{\Phi}}{F_\tau}\left(\alpha_\tau(\rho_\tau - 2) - \pi_\tau(\alpha_\tau + \rho_\tau) + S_\tau - 1\right)$$
$$\Phi_{\mathrm{QB}} \leftarrow \frac{2\widetilde{\Phi}}{F_\tau}\left(\rho_\tau(\alpha_\tau + \pi_\tau + 2) + \alpha_\tau\pi_\tau - S_\tau - 1\right)$$

12. Set the remaining conversion rates.

$$\Theta^{(1)}_{\mathrm{PA}} \leftarrow \frac{2\widetilde{\Theta}^{(1)}}{F_Y}\left(\rho_Y(\alpha_\mathrm{Y} + \pi_\mathrm{Y} + 2\alpha_\mathrm{Y}\pi_\mathrm{Y}) - \alpha_\mathrm{Y}\pi_\mathrm{Y} - S_Y + 1\right)$$
$$\Theta^{(1)}_{\mathrm{PB}} \leftarrow \frac{2\widetilde{\Theta}^{(1)}}{F_Y}\left(\pi_\mathrm{Y}(\rho_\mathrm{Y} - 2) - \alpha_\mathrm{Y}(\pi_\mathrm{Y} + \rho_Y) + S_Y - 1\right)$$
$$\Theta^{(1)}_{\mathrm{QA}} \leftarrow \frac{2\widetilde{\Theta}^{(1)}}{F_Y}\left(\alpha_\mathrm{Y}(\rho_\mathrm{Y} - 2) - \pi_\mathrm{Y}(\alpha_\mathrm{Y} + \rho_\mathrm{Y}) + S_Y - 1\right)$$
$$\Theta^{(1)}_{\mathrm{QB}} \leftarrow \frac{2\widetilde{\Theta}^{(1)}}{F_Y}\left(\rho_\mathrm{Y}(\alpha_\mathrm{Y} + \pi_\mathrm{Y} + 2) + \alpha_\mathrm{Y}\pi_\mathrm{Y} - S_Y - 1\right)$$

- For each audience, simulate user-level data.

13. Sample $Y^{(1)}_{\mathrm{Z}} \sim \mathrm{Bernoulli}\left(\Theta^{(1)}_{\mathrm{XZ}}\right)$ for all users and all Z, conditional on user type X_i.
14. Sample $Y^{(0)}_i \sim \mathrm{Bernoulli}\left(\Theta^{(0)}_{\mathrm{X}}\right)$ for all users, conditional on user type X_i.
15. Sample $\tau \sim \mathrm{Bernoulli}(\Phi_{\mathrm{XZ}})$ for all users, conditional on user type X_i and assigned ad Z_i.
16. Sample $R \sim \mathrm{Bernoulli}(\mathbf{P}(R = 1))$ for all users.

- After using τ and R to segment the audience into targeted and non-targeted sets, and the targeted users into treatment and holdout arms, compute the following:

17. For ads A and B, compute the values an advertiser would see in a typical report: $\bar{Y}^{(1)}_{\mathrm{Z,Trt}}$, $\bar{Y}^{(0)}_{\mathrm{Z,Hold}}$, $\hat{\lambda}^{\mathrm{Targ}}_{\mathrm{Z}}$, and $\widehat{\Delta}^{\mathrm{Targ}}_{\mathrm{AB}}$ (Sec. 3.3).
18. For ads A and B, compute "true" values for the audience that would not be observed directly, but are available to us for a simulated audience: $\bar{Y}^{(1)}_{\mathrm{Z,ATE}}$, $\bar{Y}^{(0)}_{\mathrm{Z,ATE}}$, $\lambda^{\mathrm{ATE}}_{\mathrm{Z}}$, and $\Delta^{\mathrm{ATE}}_{\mathrm{AB}}$.
19. Compute the bias for the audience: $\widehat{\Delta}^{\mathrm{Targ}}_{\mathrm{AB}} - \Delta^{\mathrm{ATE}}_{\mathrm{AB}}$.

For each of these conditions, we simulated 200 audience profiles using the algorithm above. Using the $\Theta^{(0)}_{\mathrm{XZ}}$, $\Theta^{(1)}_{\mathrm{XZ}}$, and Φ_{XZ} from each audience profile, and setting $\mathbf{P}(R = 1) = 0.8$, we simulated 25 audiences ($9 \times 200 \times 25 = 45,000$

audiences in total), each with 300,000 users. As in a typical report of results from a split lift test the advertiser would receive from the platform, we aggregated user outcomes across latent types, and computed the quantities in Sec. 4.2.1 for each ad and for the difference between ads. We then computed the means of those quantities across audiences with the same profile, leaving one value for each quantity for each of the 1,800 audience profiles. Each circle in Figs. 7 to 11 represents this mean value.