The A-B Test Deception: Divergent Delivery, 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

February 24, 2022

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, experimental design, A-B testing, measuring advertising effectiveness, causal inference, Simpson's paradox, social media

Consider an advertiser trying to learn about consumer preferences by comparing responses to different ads in an online field experiment. That advertiser could be an academic researching consumer behavior for scientific inquiry, a brand manager considering copy for a multichannel ad campaign, or any entity interested in learning about how customers might respond to different *creative elements* of ads (e.g., copy, images, and message). The common goal of these kinds of studies is to understand whether (and why) the content of one ad works better than another ad on the same group of users. Utilizing experimentation tools provided by online targeted advertising platforms (e.g., Google, Meta/Facebook, and Twitter) is increasingly common for this kind of consumer research. Orazi and Johnston (2020) demonstrate how to use Facebook's platform to run an online experiment that operationalizes consumer psychology theory about response to ads for COVID face masks. Experiments by Kupor et al. (2015) and Kupor and Laurin (2020) manipulate creative elements of ads on Facebook, and Cecere et al. (2018) randomize users to see different ads using Snapchat. Matz et al. (2017) features an ad experiment run through Facebook's platform aiming to test consumer responses to psychological constructs operationalized with different ad copy.

Achieving this experimental objective requires randomization, specifically to isolate the ad content's role in driving the observed outcomes from the effect of a platform's targeting algorithm placing the "right ad" in front of the "right users."[1] In a typical randomized experiment, the researcher controls the process of how subjects are assigned to treatment conditions. If that process is truly random, then the mix of characteristics of subjects exposed to one condition, on average, should be the same as the mix exposed to any other. Consumer researchers demand this level of causal inference that randomized experimentation provides. **But experiments conducted on online targeted advertising platforms are not like other experiments.** Because online experimentation tools are provided and managed by the platforms themselves, advertisers relinquish control over the most important aspects of experimental design. While the advertiser may specify that their ads in the test

[1]We are not considering a different goal of experimentation where the advertiser is actually interested in the combined effect of ad content and ad targeting. This would be the case when a marketing simply wants to know which ad "does better" on a specific platform, using all of that platform's services in combination. Ad targeting would be one of those services.

be delivered to an audience with certain characteristics (e.g., "men over 60 in Florida who are interested in gardening"), the platform controls not only which specific users from that audience to include in the test, but also how those subjects are assigned to different experimental conditions: which ads they see (Johnson 2021).

The problem with surrendering experimental control to these platforms in targeted ad environments is that the selection and assignment of users to ads may not be random at all. Not only will the targeting algorithm likely create a subject pool that is not representative of the advertiser's population of interest, the mix of subjects seeing one ad may be systematically different from the mix of another group of subjects seeing another ad. These imbalances occur because the algorithm selects subjects based on their predicted responses to the content of the ads being tested. Johnson (2021) describes this latter phenomenon as *divergent delivery* (the term we will use in this paper), while Ali et al. (2019) calls it *skewed delivery*. Divergent delivery threatens internal validity because the difference in average outcomes between ads is no longer an "apples-to-apples" comparisons. In a two-page comment, Eckles et al. (2018) illustrate this effect by showing how comparisons across different ads in Matz et al. (2017) are actually made between groups of targeted users with different demographic profiles.

Eckles et al. were able to recognize divergent delivery in the Matz et al. (2017) data because the platform broke down experimental outcomes along *observable* dimensions. But targeting does not occur solely on the basis of coarsely identifiable demographic groups. In fact, the entire mechanism and set of criteria the platform uses to decide which users see which ads is unobservable and actively withheld from the advertiser. The targeting algorithm is, after all, valuable intellectual property of the publisher, and forms a critical component of the value proposition offered to the advertiser. This means that the results the advertiser receives from the platform are aggregated across the same *unobservable* criteria the platform uses to make the targeting decisions. The treatment is confounded with the selection of subjects in unobserved ways, so attributing causal effects to the creative content of the ads themselves is not possible using the typical outcome metrics these platforms' ad experimentation tools provide. Therefore, comparing outcomes of

Ad A to ad B — the so-called A-B test— is neither a direct comparison of the creative content of A to B, nor a randomized test to measure those differences.[2] When comparing two ads, the inability to separate the ad effect from the targeting effect generates a bias between the reported A-B comparison returned by platforms' ad experimentation tools and what a randomized A-B comparison should be across the advertiser's audience of interest. In practice, the advertiser can neither observe nor correct for this bias.

This paper has two complementary objectives.

Objective 1: To present a conceptual model of online ad experimentation in targeted advertising environments, from the advertiser's perspective. We introduce our conceptual framework by describing experimental outcomes in terms of potential outcomes, using language and notation similar to the familiar Rubin model of causal inference (Rubin 1974; Angrist et al. 1996). Using this framework, we illustrate features of two common designs of online experiments, show how different designs lead to inferences of different effects for different groups of users, and explain how neither resolves the problem of biased A-B comparisons. This background sets up mathematical definitions of lift, targeting policies, bias, and heterogeneity. A novel aspect of this framework is in how we describe user response heterogeneity, targeting policies and divergent delivery as probabilities and odds ratios. This approach provides a common language to study the roles the various elements of the targeted ad experimentation environment play in distorting inferences from A-B tests. Our framework diverges from other applications of the potential outcomes model to online advertising like Gordon et al. (2019) in that we treat random assignment and exposure as separate processes, yielding an array of potential outcomes that accommodates the complexity of including two or more treatment ads in a targeted ad experiment. The model helps explain why using a "placebo control ad" (say, a public service ad unrelated to the behavioral construct being studied) is not an acceptable control condition when conducting experiments on targeted ad platforms.

[2]Experiments may contain more than two ads, but to maintain focus on pairwise comparisons, we will use the term A-B test throughout. Also, we emphasize that an A-B test always involves a comparison between two ads part of the same campaign, and not, say, between a treatment ad and a "ghost ad," as in Johnson et al. (2017a) and Gordon et al. (2019).

Objective 2: To describe how the interplay among platform targeting policies, user response heterogeneity, and aggregation of data across unobserved dimensions induces advertisers to make erroneous inferences about the relative effectiveness of elements of targeted ad content. We demonstrate how different targeting policies and patterns of heterogeneity lead to different signs and magnitudes of bias. Targeting policies cause the unobserved composition of the set of *targeted* users to deviate from the composition of the audience of interest, and they cause the compositions of targeted users exposed to different ads in the experiment to deviate from each other. Whether the changes in this mix create bias in *A*-*B* test results, in what direction, and by how much, all depend on how heterogeneous user responses are to different ads. The interactions among these forces are complex, and have yet to be fully studied.

Our analysis and simulation show how divergent delivery and response heterogeneity collude to induce bias in *A*-*B* comparisons of lift among targeted users that mislead the advertiser in the interpretation of their results. For instance, divergent delivery generates more bias when there is extensive user heterogeneity, with the targeting policies favoring users with stronger responses. We show that being unaware of these patterns can have practical consequences. The bias can be so severe that it creates an undetectable *Simpson's reversal*, where the estimated direction of the effect of interest is incorrect (i.e., when ad *A* is more effective than ad *B* for all types of users, but the aggregated experimental data leads the advertiser to infer *B* to be better than *A*).[3] Further, we show that bias can accrue *even without divergent delivery* when the experimental subject pool does not reflect the intended audience (equally unrepresentative across ads), as long as users are sufficiently heterogeneous in their preferences for different ads.

Topics like online ad targeting, user heterogeneity, and covariate imbalance have been discussed separately in the extant literature, but this is the first research to unite them with formal analysis. So with the roadmap for our paper in place, we want to be clear about what this paper is not about.

- This paper is not about demonstrating that ad targeting occurs. No advertiser should be sur-

[3] Blyth (1972) describes 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).

prised that ads are targeted to users based in part on the likelihood the user will respond to the ad. Ali et al. (2019) and Eckles et al. (2018) provide examples of experimental ad treatments being targeting to specific user types, even during the course of an supposedly randomized experiment. This targeting is the antecedent of the bias we discuss in this paper.

- The literature on experimental design has long established that when the mix of characteristics of experimental subjects differs across treatments, and outcomes for each treatment are aggregated across those characteristics, both external and internal validity of comparisons between treatments are at risk (Rosenbaum and Rubin 1983). Eckles et al. (2018), Gordon et al. (2019), and Johnson (2021) all express these concerns about "covariate imbalance" in online experiments to some extent, making a similar point: controlling for the differences with only the provided observed covariates is insufficient, yet no better experimental options are available. We also acknowledge that the designers of online experimentation platforms are aware of the potential pitfalls of conducting *A*-*B* comparisons in targeted advertising environments.

- We are addressing a different problem than the problem resolved by recent work on randomized "lift studies" or "holdout tests." To be clear, we define an *A*-*B* test specifically as a comparison between two or more ads (not ad exposure and no exposure). Johnson et al. (2017a), Johnson et al. (2017b), and Gordon et al. (2019) have all proposed solutions for testing the effect of exposure to one particular ad (versus not seeing that ad) among users targeted with that ad, which is more accurately describe as a comparison between "*A*" and "not *A*" rather than between *A* and *B*. Ad platforms already allow for random holdout in their testing tools (e.g., "ghost ads" at Google and "lift tests" at Facebook), often for multiple ads at the same time. But the results of those tests are still not comparable across ads.

Instead, the focus of this paper is on characterizing the bias that results from targeting treatment ads to users during the course of an online ad experiment. We are primarily interested in the ways that ad-level targeting biases the limited aggregated results the advertiser receives from a platform conducting an experiment on their behalf. Instead of just positing the existence or nature

of targeting with divergent delivery, we quantify that bias induced by the interactions among these factors on the advertiser's observed casual effect sizes between any two ads.

An important distinguishing characteristic of our paper is that we take the perspective of the advertiser, rather than the platform or publisher. This approach is fundamentally different from nearly all extant research on online ad experiments, which is generally concerned with how a platform might design internal-use experiments or advertiser-facing tools. The publisher's perspective currently dominates the relevant literature (Bakshy et al. 2014; Johnson et al. 2017a; Johnson et al. 2017b; Eckles et al. 2018; Gordon et al. 2019; Gordon et al. 2021), most of which was research conducted in collaboration with or sponsored by the online publishers. This distinction between the advertiser's and publisher's points of view is important because the incentives of the two parties are not always aligned even if there is an exchange of advertising expenditures. One of the reasons publishers provide experimentation tools to advertisers is to demonstrate how the targeting algorithm maximizes advertiser profit by operating differently on each ad, so both parties can make more money from campaigns run after the experiment. But during an experiment, the types of advertisers we described earlier have a goal of learning about the relative effectiveness of creative elements of an ad, isolated from how the targeting algorithm acts on those elements. The advertiser alone incurs the cost of biased inferences caused by various targeting policies. In fact, the publisher may prefer the advertiser to *not* extract generalizable inferences about different ads that might be useful for other advertising channels, either online or offline, as it would facilitate comparisons and substitution of ad spend across channels. Because the interests of publisher and advertiser in de-biasing results are in conflict, the advertiser cannot expect publishers to offer a solution, without other changes to the current landscape.

Online Ad Experiments and Causal Inference

To begin, we define a set of terms, concepts, and experimental designs to describe the targeted online advertising testing problem. Advertisers run *campaigns* on *platforms* owned by *publishers*. A campaign involves n_Z *ads*, labeled $Z \in \{z_1, z_2, \ldots, z_{n_Z}\}$, where each ad is a bundle of creative ele-

ments, such as message, copy, and imagery.[4] When initializing a campaign, the advertiser defines an *audience* of users to whom the platform may deliver ads.[5] The user eventually generates an observed *outcome* $Y_i^{(\text{obs})}$, such as ad clicks, page views, or, as in our simulation, 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 $Y_i^{(\text{obs})}$. The platform runs the experiment on behalf of the advertiser by randomly assigning *every user in the audience* to exactly one ad treatment Z_i, with assignment probabilities $\boldsymbol{\zeta} = \{\zeta_{z_1}, \ldots, \zeta_{z_{n_Z}}\}$, as the *ad treatment*.[6] 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.[7] 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_i will not necessarily be exposed to ad Z_i, the user will be exposed to ad Z_i if the user is exposed to the campaign at all.

Defining Lifts and Effects in Terms of Potential Outcomes

Following the Rubin (1974) model of causal inference, 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

[4]The individual creative elements could be defined as a set of ad attributes, but we are not studying those attributes explicitly, because mathematically their coefficients are also differences in average outcomes of treatment groups.

[5]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.

[6]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.

[7]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.

Table 1: Definitions of Potential Outcomes Depend on Orthogonal Assignment and Exposure States

Assignment:	$Z_i = A$	$Z_i = B$	$Z_i = C$	$\cdots$
Exposed ($D_i = 1$)	$Y_{\text{A}}^{(1)}$	$Y_{\text{B}}^{(1)}$	$Y_{\text{C}}^{(1)}$	$\cdots$
Not exposed ($D_i = 0$)	$Y_{\text{A}}^{(0)}$	$Y_{\text{B}}^{(0)}$	$Y_{\text{C}}^{(0)}$	$\cdots$

function of an ad treatment Z and exposure state D, but the exact function is unknown, heterogeneous, and non-stationary.[8] For instance, for $n_Z = 3$ and $Z \in \{z_1 = A, z_2 = B, z_3 = C\}$, there are 6 possible states, as in Table 1. One, $Y_{\text{A}}^{(1)}$ 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, $Y_{\text{B}}^{(0)}$, is the potential outcome *if* the user were assigned to ad B, but were not exposed to that ad ($Z = B, D = 0$). While the user is endowed with all $2n_Z$ potential outcomes, the user will end up in exactly one of the $2n_Z$ possible (D, Z) states. We refer to the only potential outcome that is ever realized as $Y_i^{(\text{obs})}$; 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.

- $Y_{\text{Z}}^{(1)} - Y_{\text{Z}}^{(0)}$ is the user-level 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 ("Aud"), lift is akin to an average treatment effect. When lift is defined only among users who were actually exposed to the ad ("Exp"), it is akin to the average treatment effect on the treated.

$$\lambda_{\text{Z}}^{\text{Aud}} = \mathbf{E}\left[Y_{\text{Z}}^{(1)} - Y_{\text{Z}}^{(0)} \right] \tag{1}$$

$$\lambda_{\text{Z}}^{\text{Exp}} = \mathbf{E}\left[Y_{\text{Z}}^{(1)} - Y_{\text{Z}}^{(0)} \,\middle|\, D = 1 \right] \tag{2}$$

- $Y_{\text{A}}^{(1)} - Y_{\text{B}}^{(1)}$ 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

[8] We suppress the i subscript in $Y_{\text{Z}}^{(D)}$ to reduce notational clutter.

assigned and exposed to ad B. An analogous effect $Y_{\mathrm{A}}^{(0)} - Y_{\mathrm{B}}^{(0)}$ is the same difference in potential outcomes, but if users were *unexposed* to each ad.

- $\left(Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{A}}^{(0)}\right) - \left(Y_{\mathrm{B}}^{(1)} - Y_{\mathrm{B}}^{(0)}\right)$ 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 difference-in-differences, 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 ($\Delta_{\mathrm{AB}}^{\mathrm{Aud}}$), or different subsets of users, such as for only exposed users ($\Delta_{\mathrm{AB}}^{\mathrm{Exp}}$).

$$\Delta_{\mathrm{AB}}^{\mathrm{Aud}} = \lambda_{\mathrm{A}}^{\mathrm{Aud}} - \lambda_{\mathrm{B}}^{\mathrm{Aud}} = \mathbf{E}\Big[\left(Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{A}}^{(0)}\right) - \left(Y_{\mathrm{B}}^{(1)} - Y_{\mathrm{B}}^{(0)}\right)\Big] \tag{3}$$

$$\Delta_{\mathrm{AB}}^{\mathrm{Exp}} = \lambda_{\mathrm{A}}^{\mathrm{Exp}} - \lambda_{\mathrm{B}}^{\mathrm{Exp}} = \mathbf{E}\Big[\left(Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{A}}^{(0)}\right) - \left(Y_{\mathrm{B}}^{(1)} - Y_{\mathrm{B}}^{(0)}\right) \Big| D = 1\Big] \tag{4}$$

The expected *unexposed* potential outcomes, $\mathbf{E}\left[Y_{\mathrm{Z}}^{(0)}\right]$, 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 expected unexposed potential outcomes do not depend on the user's assigned ad: $\mathbf{E}\left[Y_{\mathrm{A}}^{(0)}\right] = \mathbf{E}\left[Y_{\mathrm{B}}^{(0)}\right]$. 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 Z would be the same for the users who were actually exposed, for those who were actually not exposed, and for the audience: $\mathbf{E}\left[\, Y_{\mathrm{Z}}^{(1)} \mid D = 1 \,\right] = \mathbf{E}\left[\, Y_{\mathrm{Z}}^{(1)} \mid D = 0 \,\right] = \mathbf{E}\left[Y_{\mathrm{Z}}^{(1)}\right]$. Similarly, for the other exposed potential outcome, $\mathbf{E}\left[\, Y_{\mathrm{Z}}^{(0)} \mid D = 1 \,\right] = \mathbf{E}\left[\, Y_{\mathrm{Z}}^{(0)} \mid D = 0 \,\right] = \mathbf{E}\left[Y_{\mathrm{Z}}^{(0)}\right]$. Therefore, *in this case of randomized exposure*, $\lambda_{\mathrm{Z}}^{\mathrm{Exp}} = \lambda_{\mathrm{Z}}^{\mathrm{Aud}}$. Now, if we combine the random ad assignment and random exposure conditions, then $\mathbf{E}\left[\, Y_{\mathrm{A}}^{(0)} \mid D = 1 \,\right] = \mathbf{E}\left[\, Y_{\mathrm{B}}^{(0)} \mid D = 1 \,\right]$. So again, in this special case of random exposure, the A-B difference reduces to a difference in only the *exposed* potential outcomes: $\Delta_{\mathrm{AB}}^{\mathrm{Exp}} = \mathbf{E}\left[Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{B}}^{(1)} \,\middle|\, D = 1\right] = \mathbf{E}\left[Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{B}}^{(1)}\right] = \Delta_{\mathrm{AB}}^{\mathrm{Aud}}$. When initial ad assignment is randomized, but exposure is not, then $\mathbf{E}\left[\, Y_{\mathrm{A}}^{(0)} \mid D = 1 \,\right]$ and $\mathbf{E}\left[\, Y_{\mathrm{B}}^{(0)} \mid D = 1 \,\right]$ are not necessarily

equal, and differences between them would be attributable only to selection of different mixes of users, even though both groups are unexposed to either ad.

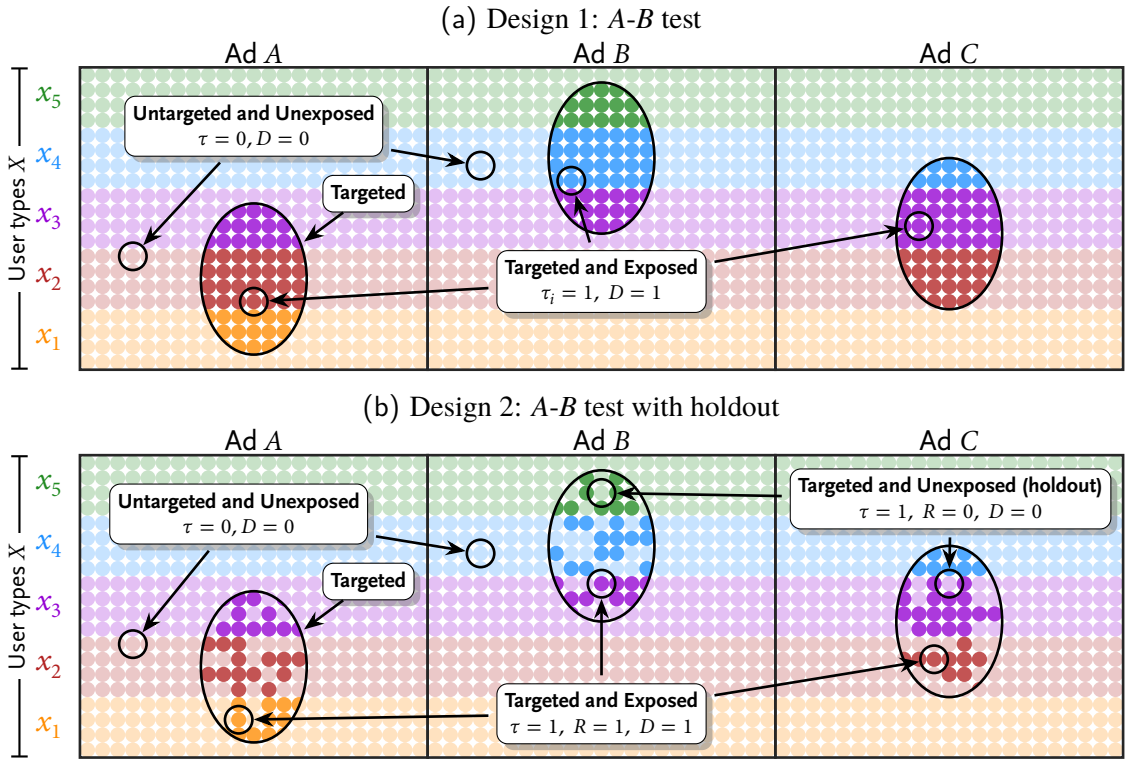
To fully appreciate how targeting affects estimated effects, we need to think about assignment as a completely different concept from exposure. Because we want to separate the effect of exposure to an ad (e.g., $\lambda_{\mathrm{A}} = \mathbf{E}\left[Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{A}}^{(0)}\right]$, a within-ad difference) from the effect of the initial ad assignment to an ad (a between-ad difference, or interaction like $\Delta_{\mathrm{AB}} = \lambda_{\mathrm{A}} - \lambda_{\mathrm{B}}$) , we need access to all $2n_Z$ potential outcomes. 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.

Targeting, Effect Estimation, and Experimental Design

The realities of targeting policies, heterogeneity, and experimental design dictate which causal effects can be estimated from the data. 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). Figure 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 "audience-ad squares," each corresponding to an ad. 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 conceptualize the targeting algorithm as a function, $\tau_{Z_i} = \tau(Z_i, X_i, \Omega) \in \{0, 1\}$. For example, for a user randomly assigned to ad $Z = A$, if they are targeted, then $\tau_A = 1$, and if not, then $\tau_A = 0$.

Figure 1: *A-B* test and *A-B* test with holdout Designs for Online Ad Experiments



Note: 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* test (Fig. 1a), all targeted users are exposed, while in a *A-B* test with holdout (Fig. 1b), users are randomly assigned to a treatment arm or holdout arm.

With τ_{Z_i}, we emphasize the dependence of the targeting decision on the ad to which the user was randomly assigned, which is a focal point of the experimental design concerns. The generic placeholder Ω contains any other information the algorithm has at its disposal from across the platform, such as other users' types and histories, diurnal or seasonal timing of the ad, the state of the auction (e.g., competition, bids), and parameters of the campaign. 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 test. The first design with these features is the *A*-*B* test (Fig. 1a). Intended for comparisons across distinct ad creatives, the *A*-*B* test design lets the advertiser compare outcomes from users assigned, targeted, and exposed to ad *A* to outcomes from users assigned, targeted, and exposed to ads *B* or *C*. In this design, all targeted users are exposed to their assigned ads (meaning that $\tau_Z = D$), 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 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 set up of Design 1 presents two immediate and practical concerns regarding interpretation of comparisons across ads in *A*-*B* tests.

Concern 1: Targeted vs untargeted mixes. In an *A*-*B* 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 (in-

side 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: $\mathbf{E}\left[Y_{\mathrm{Z}}^{(0)} \mid \tau = 1 \right] \neq \mathbf{E}\left[Y_{\mathrm{Z}}^{(0)} \mid \tau = 0 \right]$. 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 $\mathbf{E}\left[Y_{\mathrm{Z}}^{(1)} \,\middle|\, \tau = 1 \right]$ and $\mathbf{E}\left[Y_{\mathrm{Z}}^{(0)} \,\middle|\, \tau = 0 \right]$ are collected from sets of targeted and untargeted users with different mixtures of types, the advertiser cannot know if an estimate of $\mathbf{E}\left[Y_{\mathrm{Z}}^{(1)} \,\middle|\, \tau = 1 \right] - \mathbf{E}\left[Y_{\mathrm{Z}}^{(0)} \,\middle|\, \tau = 0 \right]$ 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 each ad for the audience.

Concern 2: A vs B mixes. The second concern with the *A*-*B* 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* (vertical positions of the targeting ovals in Fig. 1a). We have $\tau_A \neq \tau_B$, so the mix of users targeted with *A* and the mix targeted with *B* will not necessarily be equivalent mixes, as would be the case if the targeting did not *diverge* across experimental treatments. As a result, ad *A*'s distribution of potential outcomes among targeted users is not the same as *B*'s. While the estimates of the quantities $\mathbf{E}\left[Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{A}}^{(0)} \,\middle|\, \tau = 1 \right]$ and $\mathbf{E}\left[Y_{\mathrm{B}}^{(1)} - Y_{\mathrm{B}}^{(0)} \,\middle|\, \tau = 1 \right]$ can *separately* be interpreted as causal effects, the difference in these lifts, $\mathbf{E}\left[Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{A}}^{(0)} \,\middle|\, \tau = 1 \right] - \mathbf{E}\left[Y_{\mathrm{B}}^{(1)} - Y_{\mathrm{B}}^{(0)} \,\middle|\, \tau = 1 \right]$, 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.[9]

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 ad fails to reveal a true causal effect of exposure to that ad. When

[9]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.

making this comparison in a standard randomized experiment, 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 ($Y_C^{(1)} = Y_A^{(0)}$), to make $\mathbf{E}\left[Y_A^{(1)} - Y_C^{(1)} \,\middle|\, \tau = 1\right] = \mathbf{E}\left[Y_A^{(1)} - Y_A^{(0)} \,\middle|\, \tau = 1\right]$. 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: the A-B test with holdout. Fig. 1b illustrates a A-B test with holdout design, sometimes known as a "split lift test." This design introduces a distinction between exposure and targeting, which are equivalent in Design 1. 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).

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 ad that would have been shown to user i in that exact time and context, if Z_i did not exist (i.e., the second place ad in the auction). Among the targeted users, the advertiser only observes results for each ad that are aggregated by treatment arm ($\bar{Y}^{(1)}_{Z,\text{Exp}}$) and holdout arm ($\bar{Y}^{(0)}_{Z,\text{Hold}}$). We consider a user who is assigned to Z_i to be "in compliance" if that user is also exposed to Z_i, where compliance depends on two distinct factors: the targeting mechanism indicated by τ, and the random holdout mechanism indicated by R.

Concern 1 is partially resolved by Design 2. Design 2's additional randomization step among targeted users is a recent innovation in online experimentation. 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 has been 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 counterfactual comparison of potential outcomes that defines a single ad's lift.

Platforms that implement this design are essentially running "two-armed mini randomized experiments" among only the targeted users (Gordon et al. 2019), but each experiment is run on a group that is targeted differently for each ad. 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 $\mathbf{E}\left[Y_{\mathrm{Z}}^{(1)} - Y_{\mathrm{Z}}^{(0)} \,\middle|\, \tau = 1\right]$ does have a causal interpretation for the set of targeted users. This design is analogous to a design where the targeting process is flagging users who are "intended to be treated" (ITT). But when targeting is non-random (mixes of colors in the targeting oval are different from the mixes in the entire audience square), then $\mathbf{E}\left[Y_{\mathrm{Z}}^{(1)} - Y_{\mathrm{Z}}^{(0)} \,\middle|\, \tau = 1\right] \neq \mathbf{E}\left[Y_{\mathrm{Z}}^{(1)} - Y_{\mathrm{Z}}^{(0)}\right]$. Therefore, we cannot say that Concern 1 is resolved completely.

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 randomized experiment, where

$$\mathbf{E}\left[Y_{\mathrm{Z}}^{(1)} - Y_{\mathrm{Z}}^{(0)} \,\middle|\, \tau_Z = 1\right] = \mathbf{E}\left[Y_{\mathrm{Z}}^{(1)} \,\middle|\, R = 1, \tau_Z = 1\right] - \mathbf{E}\left[Y_{\mathrm{Z}}^{(0)} \,\middle|\, R = 0, \tau_Z = 1\right], \tag{5}$$

the effects are still computed from different mixes of targeted users, So even when restricting inference to the set of targeted users, the difference in lift across ads, $\mathbf{E}\left[Y_{\mathrm{A}}^{(1)} - Y_{\mathrm{A}}^{(0)} \,\middle|\, \tau_A = 1\right] - \mathbf{E}\left[Y_{\mathrm{B}}^{(1)} - Y_{\mathrm{B}}^{(0)} \,\middle|\, \tau_B = 1\right]$, does not have a causal interpretation. Even worse, *the advertiser will not see this confound and cannot detect how much of a problem it will be*.[10] 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* test with holdout (Fig. 1b), because at least it resolves part of Concern 1.

As a summary, the tree in Fig. 2 ties the assignment, targeting, and exposure processes together. It

[10] Some platforms provide results that are broken down by coarse demographic groups (e.g., gender, age). In our framework, those kinds of groups 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.

Figure 2: Conceptual Tree Framework for an *A*-*B* Test with Holdout 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.

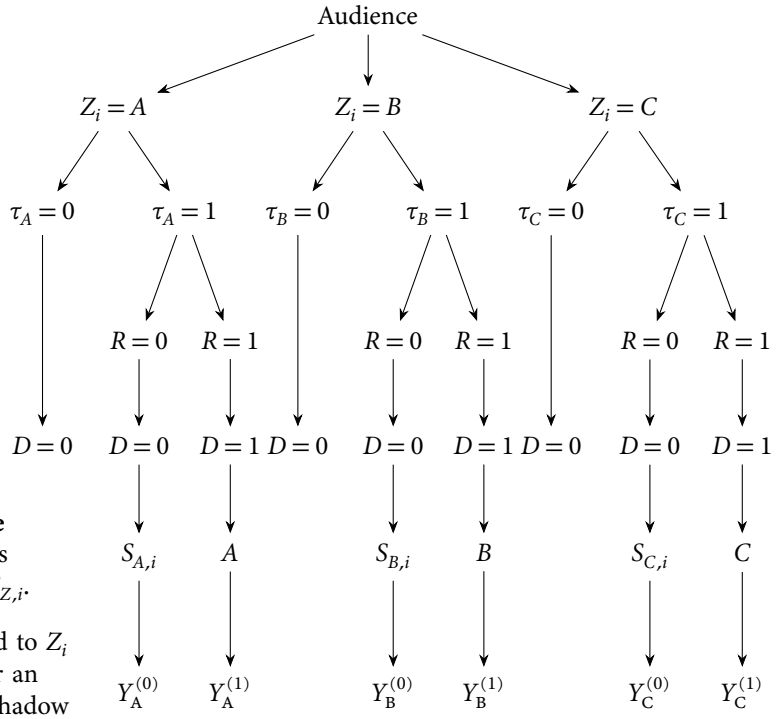
Untargeted users ($\tau_Z = 0$) will see ads that are unassociated with the focal campaign.

Targeted users ($\tau_Z = 1$) 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 $D = \tau \cdot R = 1$.

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: 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.

provides an annotated tour of how exposure to one of the experimental treatment ads requires the user to pass through the "filters" of targeting and random assignment to the treatment arm.

Defining Causal Effects with Targeting and Heterogeneity

Now, we mathematically formalize our intuition about how targeting algorithms generate problematic comparisons of effects between ads. This section extends our notation and definitions; the analytic and simulation results will follow in the subsequent section.

Lift and Targeting with Heterogeneous Users

As introduced earlier, a user type X 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:

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

The *aggregate* lift λ_Z^{Aud} (marginal across ads) and the aggregate A-B difference Δ_{AB}^{Aud} from Eqs. 1 and 3 are mixtures of their type-specific counterparts, with γ_X as the mixture weights:

$$\lambda_Z^{\text{Aud}} = \sum_{\forall X} \lambda_{XZ} \gamma_X \tag{7}$$

$$\Delta_{AB}^{\text{Aud}} = \sum_{\forall X} \gamma_X \left(\lambda_{XA} - \lambda_{XB} \right) \tag{8}$$

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

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

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.

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

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

$$\widetilde{\Phi} = \mathbf{P}(\tau = 1) = \sum_{\forall Z} \Phi_Z \zeta_Z, \tag{11}$$

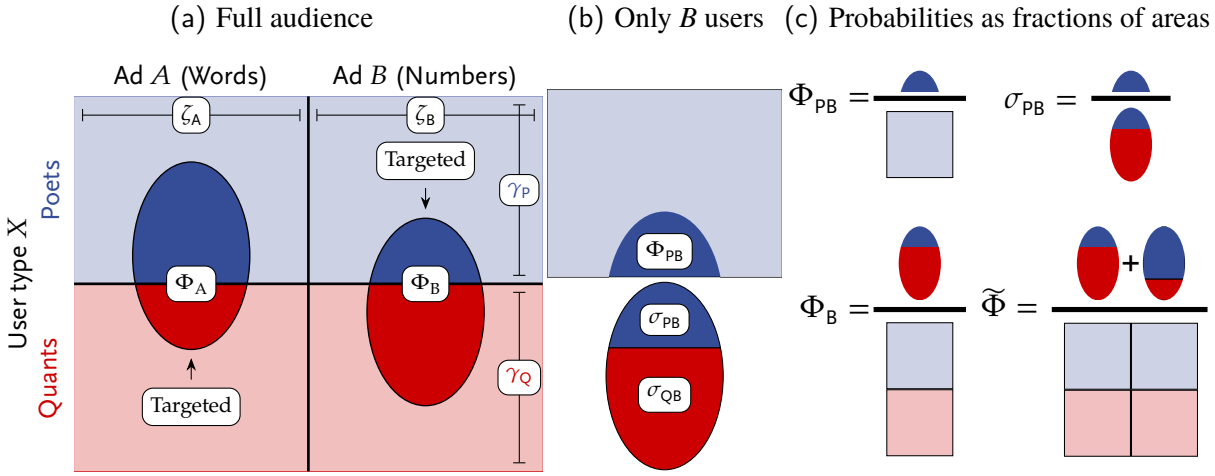
The distribution of user types among only targeted users who were assigned to ad Z is different from distribution of user types among the audience. While γ_X represents the prior mixture of user types among all users in the audience, we define σ_{XZ} to be the posterior mixture only among targeted users.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

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

To better understand the role of user type X, for the rest of this paper we will consider a world with two user types, P and Q, and two ads, A and B. For expositional clarity, we will name the user types Poets and Quants. Also, we describe ad A as a *Words* ad because its creative elements contain more verbal content, and ad B as a *Numbers* ad because it contains more quantitative content. Using these conceptual descriptions lets us characterize targeting policies where, for example, Poets might receive more ads with more words, and Quants might receive more ads with more numbers.

Figure 3 illustrates the definitions of Φ_{XZ}, Φ_X, ζ_Z, γ_X and σ_{XZ}, with a two-ad experiment and a two-type audience. In Fig. 3a, areas of the two dark "targeting ovals" in each column are the same proportions as the areas of their respective columns ($\Phi_A = \Phi_B$). But in this example, *Poets* who are randomly assigned to the Numbers ad (B) are less likely to be targeted than a *randomly chosen user* assigned to the Numbers ad ($\Phi_{PB} < \Phi_B$). Visually, the proportion of the right blue square in Fig. 3a that is inside the targeting oval (Φ_{PB}, also in the top of Fig. 3b) is smaller than the proportion of the right *column* of Fig. 3a that is inside (Φ_B). From Eq. 12, $\sigma_{PB} < \gamma_P$, so the blue proportion of the B oval in Fig. 3b (bottom) is smaller than the blue proportion of the entire audience (Fig. 3a, full grid). On the other hand, A-Poets (Poets who were randomly assigned to the Words ad A) are *more* likely to be targeted than A users overall ($\Phi_{PA} > \Phi_A$), so $\sigma_{PA} > \gamma_P$. Therefore, $\sigma_{PA} > \gamma_P > \sigma_{PB}$. We distinguish between these two effects: (1) targeting by user types overall occurs as the posterior mixture probability deviates from the prior mixture ($\sigma_{XZ} \neq \gamma_X$); and (2) *divergent delivery* 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: 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_{PB} is the probability that a B-Poet is targeted (proportion of the B-Poet quadrant inside the oval). Fig. 3b, bottom: Posterior probabilities σ_{PB} and σ_{QB} are proportions of users *targeted* with B who are Poets and Quants, respectively. Fig. 3c defines probabilities as fractions of not-to-scale areas.

when the mix of users targeted with one ad does not resemble the mix targeted with another ad ($\sigma_{XA} \neq \sigma_{XB}$).

Characterizing a Campaign's Targeting Policy

We now characterize the targeting policies that underlie the probabilities illustrated in Fig. 3 with a parsimonious set of ratios: Eqs. 13 to 15. These ratios define relationships among targeting probabilities between ads (α_τ), between user types (π_τ), and their interactions (ρ_τ). The second equality in Eq. 15 comes from substitution of each σ_{XZ} after solving Eq. 12.

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

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

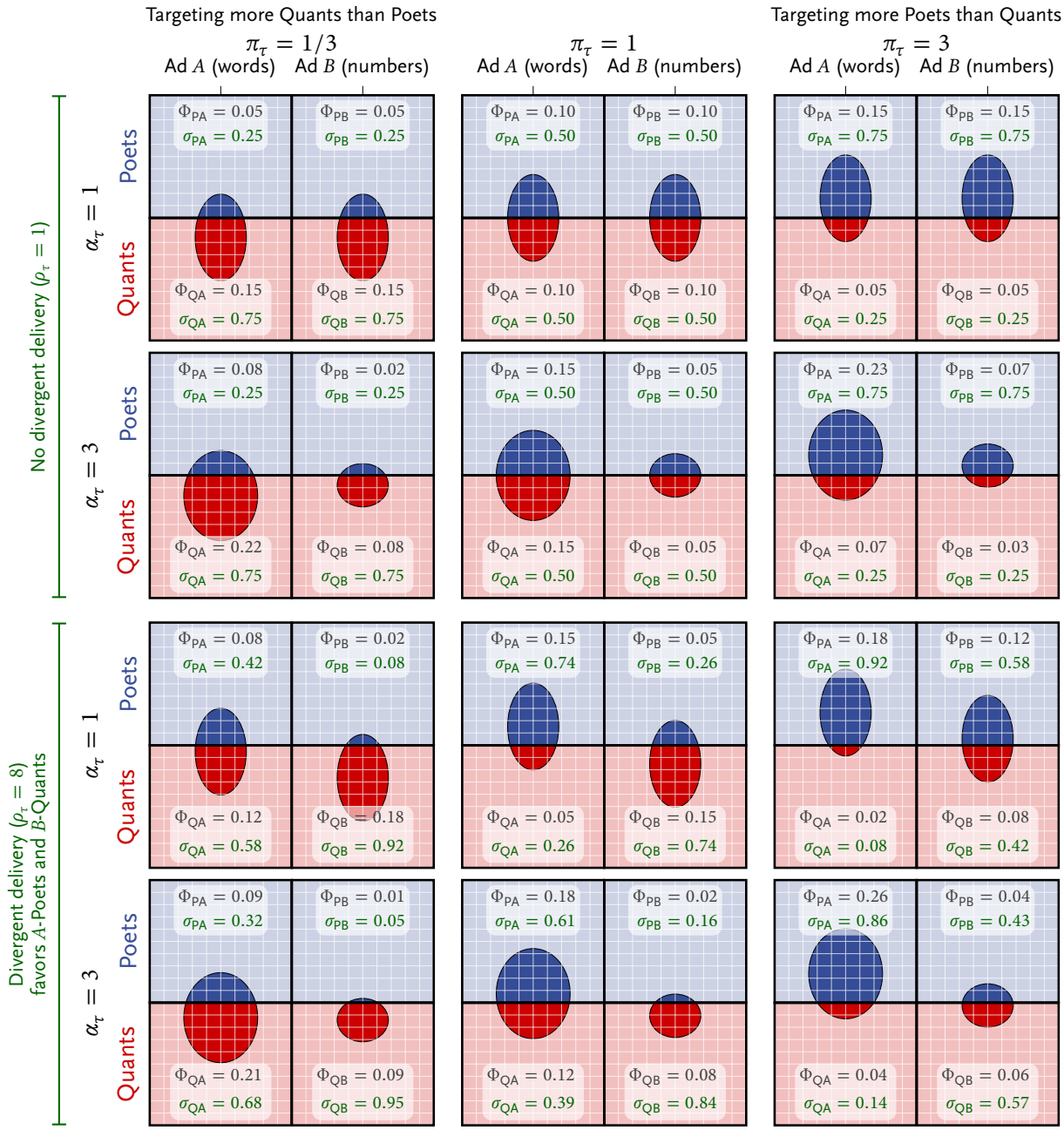
$$\rho_\tau = \frac{\Phi_{PA}}{\Phi_{PB}} \Big/ \frac{\Phi_{QA}}{\Phi_{QB}} = \frac{\sigma_{PA}}{1-\sigma_{PA}} \Big/ \frac{\sigma_{PB}}{1-\sigma_{PB}} \qquad \text{(divergent delivery)} \qquad (15)$$

These ratios answer specific questions about a platform's algorithmic targeting policies in terms of pairwise comparisons across user types and ads. 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 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 the Words ad (A) more than the Numbers ad (B). How differently does the algorithm target the Words and Numbers ads 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 algorithm favors those Poets assigned to the Words ad (A) and those Quants assigned to the Numbers ad (B) *even more than* whatever the marginal ratios α_τ and π_τ alone would have indicated. While we could 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 our 2 ads × 2 user types design.

Figure 4 illustrates how different combinations of these three ratios' values correspond to distinct targeting policies; the Web 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).

By characterizing algorithmic targeting using these ratios, we can see a relationship between divergent delivery (ρ_τ) and the posterior mixtures of targeted users (σ). We explain with two examples from Fig. 4. The top right panel describes a targeting policy that is (1) equally likely to target users assigned to the Words and Numbers ads ($\alpha_\tau = 1$; the targeting ovals have equal area); (2) is three times more likely to target Poets than Quants ($\pi_\tau = 3$; more blue area than red area across both ovals); and (3) creates a mix of targeted users that is the same across the Words and

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



Note: Each panel is an audience, divided into quadrants for each combination of ad and user type, with ad A on left, ad B on right, Poets in blue on the top, and Quants in red on the bottom. As in Fig. 3a, $\mathbf{\Phi}_{XZ}$ is the proportion of a quadrant inside a targeting oval, and σ_{XZ} is the proportion of an 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 white grid square represents 1% of the audience in a quadrant (e.g., if $\Phi_{PA} = .26$, the blue part of the A oval covers 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 targeting uses divergent delivery ("no" in the top $\rho_\tau = 1$ rows, and "yes" in the bottom $\rho_\tau = 8$ rows). In all panels, $\gamma_P = \gamma_Q = .5$, $\zeta_A = \zeta_B = .5$, and $\widetilde{\Phi} = .1$.

Numbers ads (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 $\sigma_{PA} = \sigma_{PB}$), the mix among all targeted users will not necessarily be the same as the mix in the audience ($\sigma_{PA} = \sigma_{PB} \neq \gamma_P$). The targeted mix will be the same as the audience mix only when $\rho_\tau = 1$ *and* $\pi_\tau = 1$. In the bottom center panel of Fig. 4, the algorithm is more likely to target users assigned to Words than Numbers ($\alpha_\tau = 3$; 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* Words users to be Poets and a higher proportion of the *targeted* Numbers users to be Quants ($\sigma_{PA} \neq \sigma_{PB}$; the vertical positioning of the center of the ovals is higher for *A* than *B*).

Estimation of Effects and Bias

The reason divergent delivery causes problems for causal inference in *A*-*B* comparisons is that variance in targeting probabilities across user types and ads is equivalent to separation of the posterior mixing probabilities among the targeted set of users. In practice, the problem arises because the aggregate experimental results that are reported to the advertiser, and are computed from outcomes of targeted users, do not actually reflect the effects the advertiser wants to measure.

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

$$\lambda_Z^{Targ} = \mathbf{E}\left[Y_Z^{(1)} - Y_Z^{(0)} \,\middle|\, \tau = 1 \right] = \lambda_{PZ}\sigma_{PZ} + \lambda_{QZ}\sigma_{QZ} \tag{16}$$

The *A*-*B* difference Δ_{AB}^{Targ} is a difference in lifts among users targeted with each ad, and is thus a difference-in-differences of expected potential outcomes.

$$\Delta_{AB}^{Targ} = \lambda_A^{Targ} - \lambda_B^{Targ} = \left(\lambda_{PA}\sigma_{PA} - \lambda_{PB}\sigma_{PB}\right) + \left(\lambda_{QA}\sigma_{QA} - \lambda_{QB}\sigma_{QB}\right) \tag{17}$$

Each parenthetical term in Eq. 17 is the contribution of the corresponding user 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), because targeting only affects the mix (the shift from γ_X to σ_{XZ}). The λ_{XZ} are unaffected.

In an experiment with a randomly-determined holdout sample, exposure among the targeted users is random even when targeting itself is not. Therefore, $\lambda_Z^{\text{Targ}} = \lambda_Z^{\text{Exp}}$. If targeting among the audience were entirely random ($\pi_\tau = 1$, $\rho_\tau = 1$), then $\sigma_{XA} = \sigma_{XB} = \gamma_X$. This is the only situation in which $\Delta_{AB}^{\text{Targ}} = \Delta_{AB}^{\text{Aud}}$. 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 $\sigma_{XA} \neq \sigma_{XB} \neq \gamma_X$. But these two non-random targeting cases tell two separate stories. The expected difference between the lift of *A* among users targeted with *A* and the lift of *B* among users targeted with *B* ($\rho_\tau \neq 1$) is not equivalent to the difference between lifts of *A* and *B* when targeted to identical mixes by the overall campaign ($\rho_\tau = 1$).

Estimates

The distinctions between λ_Z^{Targ} and λ_Z^{Aud}, and between $\Delta_{AB}^{\text{Targ}}$ and Δ_{AB}^{Aud}, 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* test with holdout design, 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 λ_Z^{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.

$$\widehat{\lambda}_Z^{\text{Targ}} = \bar{Y}_{Z,\text{Exp}}^{(1)} - \bar{Y}_{Z,\text{Hold}}^{(0)} \tag{18}$$

$$\widehat{\Delta}_{AB}^{\text{Targ}} = \widehat{\lambda}_A^{\text{Targ}} - \widehat{\lambda}_B^{\text{Targ}} \tag{19}$$

Equations 16 to 17 let us formalize the concerns faced by advertisers who rely on the estimators in Eqs. 18 and 19 for inferences about the effectiveness of their ad creatives:

- The advertiser will see an ad's $\widehat{\lambda}_{Z}^{Targ}$, which is an unbiased estimate for λ_{Z}^{Targ}. But λ_{Z}^{Targ} is actually a mixture of λ_{XZ} for each type, and the advertiser does not have an estimator for those unobserved user-type-specific lifts for each ad. This aggregate estimate confounds variation in λ_{XZ} with variation in σ_{XZ}. Multiplying λ_{XZ} by a constant, and dividing σ_{XZ} by that same constant, leaves λ_{Z}^{Targ} (and $\widehat{\lambda}_{Z}^{Targ}$) unchanged.[11] Thus, the advertiser cannot know how much of the observed lift is due to users' responses to the ad's creative elements themselves, or to the algorithm's method of choosing which users will receive that ad.
- If the advertiser cares only about lift of that one ad on *targeted* users, then the estimate of λ_{Z}^{Targ} 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}^{Aud} instead. Because non-random targeting means that $\sigma_{XZ} \neq \gamma_{X}$, even the true values for λ_{Z}^{Targ} are not equal to λ_{Z}^{Aud}, and so estimating $\widehat{\lambda}_{Z}^{Targ}$ does not help (Concern 1).
- If the advertiser does not care about isolating drivers of differences in responses between ads, then $\widehat{\Delta}_{AB}^{Targ}$ satisfies those needs even though Δ_{AB}^{Targ} does not equal Δ_{AB}^{Aud}. But the advertiser who wants to separate the effect of ad content from how the algorithm targets users with each ad cannot do so because $\widehat{\lambda}_{A}^{Targ}$ and $\widehat{\lambda}_{B}^{Targ}$ are computed from different mixes of users ($\sigma_{XA} \neq \sigma_{XB}$). Thus, the $\widehat{\Delta}_{AB}^{Targ}$ from *A*-*B* test with holdout does not solve this advertiser's problem when comparing two ads (Concern 2). This is the problem facing researchers who are testing behavioral hypotheses that are operationalized by ads with different creative elements.

Bias

For the rest of this paper, we will focus on how divergent delivery leads the estimated effect from the targeted users, $\widehat{\Delta}_{AB}^{Targ}$, to deviate from the effect for the audience, Δ_{AB}^{Aud}. We define another difference-in-differences, $\widehat{\mathscr{E}}_{Z}^{\lambda} = \widehat{\lambda}_{Z}^{Targ} - \lambda_{Z}^{Aud}$, to be the *bias* in the estimate of an ad's lift computed

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

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.

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

Eq. 20 is a "diff-in-diff-in-diff." First differences are between outcomes of treatment (exposed) and holdout groups: $\widehat{\lambda}^{\text{Targ}}_{Z} = \bar{Y}^{(1)}_{Z,\text{Exp}} - \bar{Y}^{(0)}_{Z,\text{Hold}}$ for targeted users, and $\lambda^{\text{Aud}}_{Z} = \mathbf{E}\left[Y^{(1)}_{Z} - Y^{(0)}_{Z}\right]$ for true audience values. Second differences are between ads: $\widehat{\Delta}^{\text{Targ}}_{AB} = \widehat{\lambda}^{\text{Targ}}_{A} - \widehat{\lambda}^{\text{Targ}}_{B}$ and $\Delta^{\text{Aud}}_{AB} = \lambda^{\text{Aud}}_{A} - \lambda^{\text{Aud}}_{B}$. Third differences are between targeted and audience values: $\widehat{\mathscr{E}}^{\Delta}_{AB} = \widehat{\Delta}^{\text{Targ}}_{AB} - \Delta^{\text{Aud}}_{AB}$. We use the term "bias" because $\widehat{\mathscr{E}}^{\Delta}_{AB}$ 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* test with holdout experiment's data, the estimate of an ad's lift for its targeted users is estimated without problem. The remaining issue contributing to $\widehat{\mathscr{E}}^{\lambda}_{Z}$ is that the targeted mix and audience mix differ: $\widehat{\lambda}^{\text{Targ}}_{Z} \neq \lambda^{\text{Aud}}_{Z}$.

To formally study the factors that are causing this bias, we continue with the case of two ads, Words and Numbers (*A* and *B*), and two user types, Poets and Quants. The progression of Eqs. 21 to 24 comprises a derivation of $\mathscr{E}^{\Delta}_{AB}$ for this 2×2 case by collecting and simplifying Eqs. 7, 8, 16, 17 and 20.

$$\lambda^{\text{Aud}}_{A} = \gamma_{P}\lambda_{PA} + (1-\gamma_{P})\,\lambda_{QA} \tag{21a}$$

$$\lambda^{\text{Targ}}_{A} = \sigma_{PA}\lambda_{PA} + (1-\sigma_{PA})\,\lambda_{QA} \tag{21b}$$

$$\widehat{\mathscr{E}}^{\lambda}_{A} = (\sigma_{PA}-\gamma_{P})\left(\lambda_{PA}-\lambda_{QA}\right) \tag{21c}$$

$$\lambda^{\text{Aud}}_{B} = \gamma_{P}\lambda_{PB} + (1-\gamma_{P})\,\lambda_{QB} \tag{21d}$$

$$\lambda^{\text{Targ}}_{B} = \sigma_{PB}\lambda_{PB} + (1-\sigma_{PB})\,\lambda_{QB} \tag{21e}$$

$$\widehat{\mathscr{E}}^{\lambda}_{B} = (\sigma_{PB}-\gamma_{P})\left(\lambda_{PB}-\lambda_{QB}\right) \tag{21f}$$

$$\Delta^{\text{Aud}}_{AB} = \gamma_{P}\left(\lambda_{PA}-\lambda_{PB}\right) + (1-\gamma_{P})\left(\lambda_{QA}-\lambda_{QB}\right) \tag{22}$$

$$\Delta^{\text{Targ}}_{AB} = \left(\sigma_{PA}\lambda_{PA} + (1-\sigma_{PA})\,\lambda_{QA}\right) - \left(\sigma_{PB}\lambda_{PB} + (1-\sigma_{PB})\,\lambda_{QB}\right) \tag{23}$$

$$\widehat{\mathscr{E}}^{\Delta}_{AB} = \underbrace{\overbrace{(\sigma_{PA}-\gamma_{P})}^{(24.A.1)}\overbrace{\left(\lambda_{PA}-\lambda_{QA}\right)}^{(24.A.2)}}_{(24.A)} - \underbrace{\overbrace{(\sigma_{PB}-\gamma_{P})}^{(24.B.1)}\overbrace{\left(\lambda_{PB}-\lambda_{QB}\right)}^{(24.B.2)}}_{(24.B)} \tag{24}$$

Equation 24 shows that bias comes from three sources:

- *The targeted mix differs from the audience.* For each ad, Factors 24.A.1 and 24.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. 24.A.2 and 24.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. 24.A and Eq. 24.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. 24, the amount of the bias depends on heterogeneity in users' responses to the two ads through the $\left(\lambda_{PA} - \lambda_{QA}\right)$ and $\left(\lambda_{QB} - \lambda_{PB}\right)$ factors. These factors are the partial derivatives of $\widehat{\mathscr{E}}^{\Delta}_{AB}$ with respect to the targeted mix of users. For example, if Poets are more responsive to the Words ad (A) than Quants, then targeting more Poets among users assigned to the Words ad will push $\widehat{\lambda}^{\text{Targ}}_{A}$ above λ^{Aud}_{A}, driving up $\widehat{\Delta}^{\text{Targ}}_{AB}$ relative to Δ^{Aud}_{AB}, and increasing the bias. But if Poets are more responsive than Quants to the Numbers ad B as well, then targeting more Poets among the B-users makes $\widehat{\lambda}^{\text{Targ}}_{B}$ larger than λ^{Aud}_{B}, driving *down* $\widehat{\Delta}^{\text{Targ}}_{AB}$ and offsetting the increase in bias from A. If Poets were less responsive than Quants to the Numbers ad, 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 Web Appendix.

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 notational clutter, define potential outcomes $\Theta^{(1)}_{XZ} = \mathbf{E}\left[\, Y^{(1)}_{Z} \mid X \,\right]$, $\Theta^{(0)}_{XZ} = \mathbf{E}\left[\, Y^{(0)}_{Z} \mid X \,\right]$, $\Theta^{(1)}_{Z} = \mathbf{E}\left[Y^{(1)}_{Z} \right]$, $\Theta^{(0)}_{Z} = \mathbf{E}\left[Y^{(0)}_{Z} \right]$, $\Theta^{(1)}_{X} = \mathbf{E}\left[\, Y^{(1)} \mid X \,\right]$, and $\Theta^{(0)}_{X} = \mathbf{E}\left[\, Y^{(0)} \mid X \,\right]$. Thus, we can write Eq. 6 as

$\lambda_{XZ} = \Theta^{(1)}_{XZ} - \Theta^{(0)}_{XZ}$ and Eq. 7 as $\lambda^{\text{Aud}}_{Z} = \Theta^{(1)}_{Z} - \Theta^{(0)}_{Z}$. 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).

$$\alpha_Y = \frac{\Theta^{(1)}_{A}}{\Theta^{(1)}_{B}} = \frac{\gamma_P \Theta^{(1)}_{PA} + \gamma_Q \Theta^{(1)}_{QA} + \cdots}{\gamma_P \Theta^{(1)}_{PB} + \gamma_Q \Theta^{(1)}_{QB} + \cdots} \qquad \text{(relative ad effectiveness)} \tag{25}$$

$$\pi_Y = \frac{\Theta^{(1)}_{P}}{\Theta^{(1)}_{Q}} = \frac{\zeta_A \Theta^{(1)}_{PA} + \zeta_B \Theta^{(1)}_{PB} + \cdots}{\zeta_A \Theta^{(1)}_{QA} + \zeta_B \Theta^{(1)}_{QB} + \cdots} \qquad \text{(user heterogeneity)} \tag{26}$$

$$\rho_Y = \frac{\Theta^{(1)}_{PA}}{\Theta^{(1)}_{PB}} \Big/ \frac{\Theta^{(1)}_{QA}}{\Theta^{(1)}_{QB}} \qquad \text{(user-ad response interaction)} \tag{27}$$

The α_Y ratio captures relative *ad effectiveness* overall. If $\alpha_Y > 1$, the expected response after exposure to the Words ad (A) is greater than the Numbers 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 the Words ad and Quants respond even better to the Numbers ad than whatever the marginal ratios α_Y and π_Y alone would have dictated. Because these ratios define the same kinds of relationships as for targeting policies, we refer the reader back to Fig. 4 for a visualization.

Bringing Targeting and Heterogeneity Together

We now tie together the various pieces of our conceptual framework:

- α_τ, π_τ, and ρ_τ describe *targeting policies* as relationships among targeting probabilities Φ_{XZ} for all user types X and ads 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 describe *response heterogeneity* in users' potential outcomes after being exposed to each ad for each user type, λ_{XZ} for all X and Z.

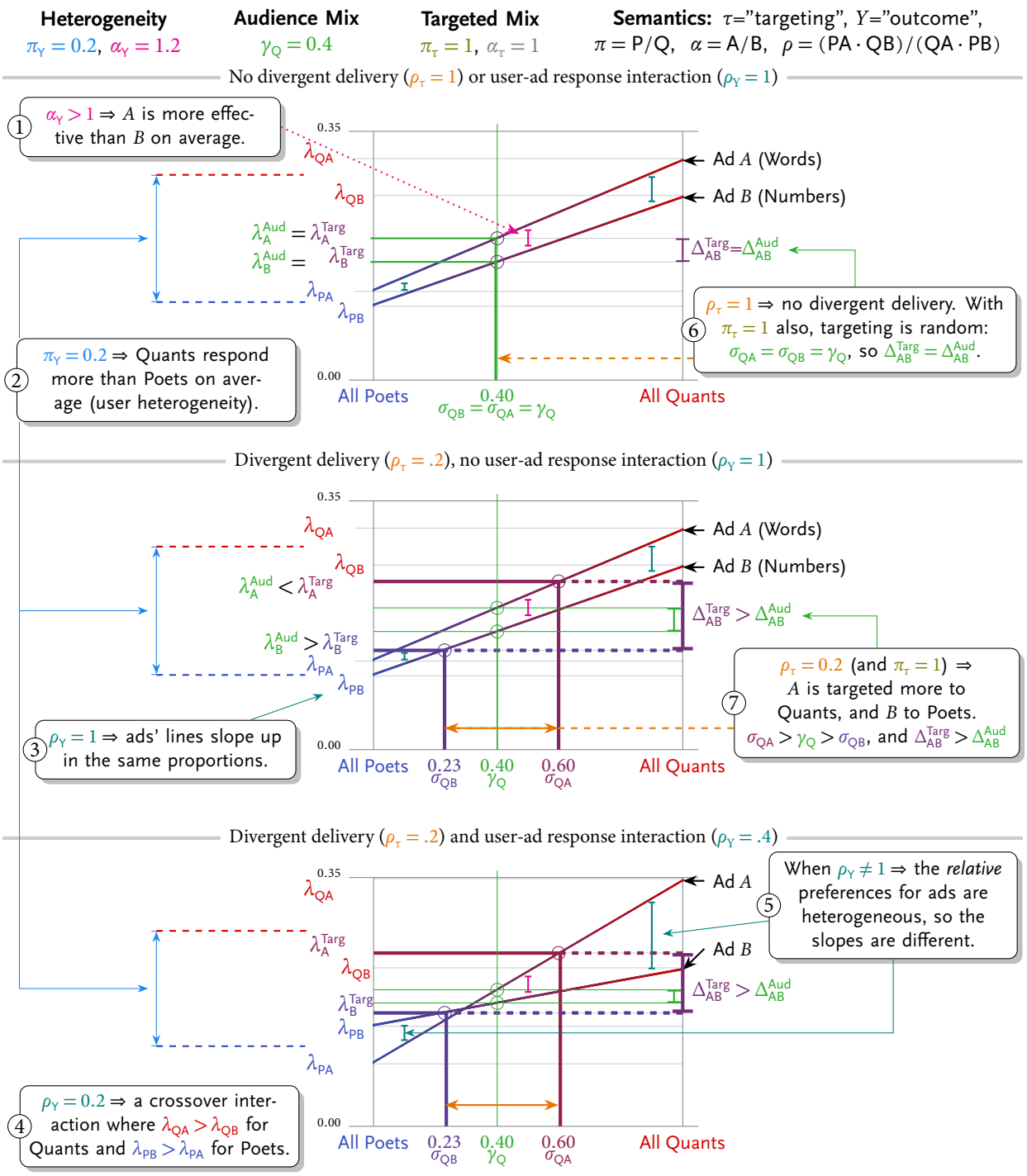
- The estimated lift for each ad, $\widehat{\lambda}_{Z}^{Targ}$, is a mixture of λ_{XZ}, with σ_{PZ} providing the mixture weights. $\widehat{\Delta}_{AB}^{Targ}$ is the difference in those estimated lifts.
- The bias in *A*-*B* difference, $\widehat{\mathscr{E}}_{AB}^{\Delta}$, is the difference-in-differences of the effects of ads between the true value for audience Δ_{AB}^{Aud} and the estimate for targeted users $\widehat{\Delta}_{AB}^{Targ}$.

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 walk through how the mix of user types determines the ad's aggregate lift. Each ad's targeted lift λ_{Z}^{Targ} (y-axis; Eqs. 21b and 21e) 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 the type-specific lift (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 ($\sigma_{PZ} = \sigma_{QZ} = \gamma_{Q}$), then aggregate lifts would be λ_{Z}^{Aud} (green lines and notes).

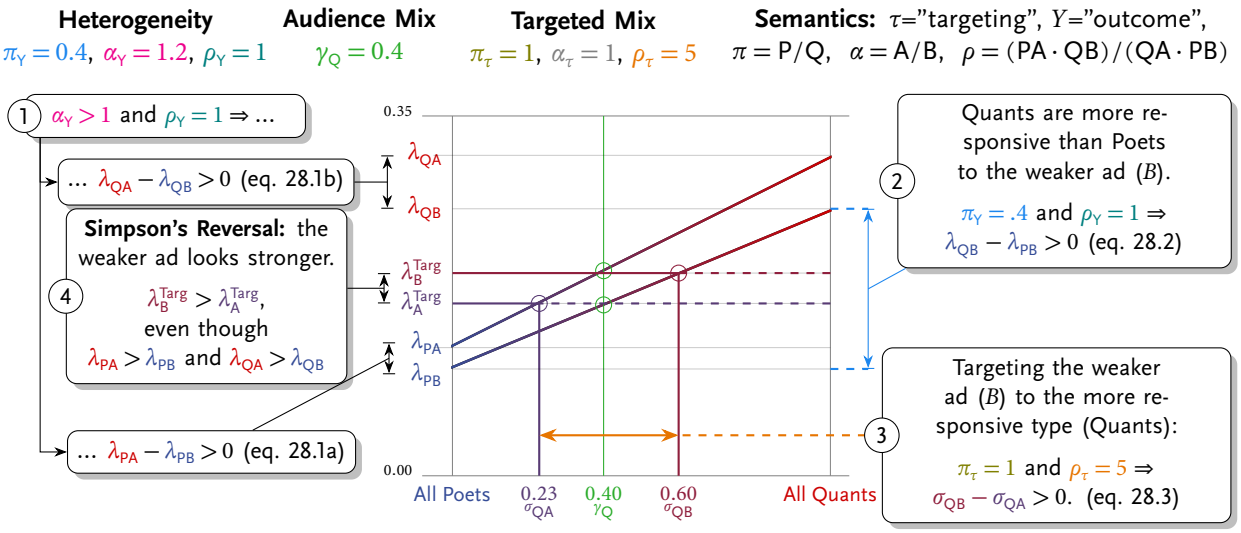
The panels in Fig. 5 differ by their levels of divergent delivery (ρ_{τ}) and user-ad response interaction (ρ_{Y}). With divergent delivery ($\rho_{\tau} = .2$, middle and bottom panels), the targeted mixes are different for each ad, so σ_{QA} and σ_{QB} deviate from γ_{Q}, and from each other (horizontal 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., $\lambda_{A}^{Targ} > \lambda_{A}^{Aud}$). 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}). Thus, the *same level of divergent delivery has different effects depending on the nature of response heterogeneity*. With no response heterogeneity ($\rho_{Y} = 1$, middle panel), $\lambda_{PA} > \lambda_{PB}$ and $\lambda_{QA} > \lambda_{QB}$ (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_{AB}^{Targ} relative to Δ_{AB}^{Aud}. In

Figure 5: Effects of Targeting and Heterogeneity on Aggregate Lift



Note: Targeted aggregate lifts λ_Z^{Targ} (y-axis) are linear combinations (purple) of λ_{PZ} and λ_{QZ} (blue and red ends, respectively), weighted by targeted (posterior) mixture probabilities of Quants, σ_{QA} and σ_{QB} (x-axis). Divergent delivery induces these weights to deviate from audience (prior) mixture probability γ_Q (green). 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), $\widetilde{\Phi} = .2$ (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. 28



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.

Simpson's Reversal

A particularly concerning example of $\widehat{\Delta}_{AB}^{Targ}$ being a poor estimate of Δ_{AB}^{Aud} 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 incorrectly show that B is stronger than A; that is, if $\lambda_{PA} > \lambda_{PB}$ and $\lambda_{QA} > \lambda_{QB}$, but $\lambda_A^{Targ} < \lambda_B^{Targ}$.

Figure 6 illustrates how a Simpson's reversal can happen. In this example, the Words ad A has a higher lift than the Numbers ad B for both user types. But Quants are so much more responsive than Poets overall that any ad's mixture in which Quants are 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 $\widehat{\lambda}_B^{Targ}$ will be too high, the estimated $\widehat{\lambda}_A^{Targ}$ will be too low, and a Simpson's reversal will occur.

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

$$\overbrace{\left(\lambda_{QB}-\lambda_{PB}\right)}^{(28.2)}\overbrace{\left(\sigma_{QB}-\sigma_{QA}\right)}^{(28.3)} > \underbrace{\left(1-\sigma_{QA}\right)\overbrace{\left(\lambda_{PA}-\lambda_{PB}\right)}^{(28.1a)}+\sigma_{QA}\overbrace{\left(\lambda_{QA}-\lambda_{QB}\right)}^{(28.1b)}}_{(28.1)} \tag{28}$$

Equation 28 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* (28.1); (2) the difference between user types for the *weaker* ad's lift is sufficiently large (28.2); 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 (28.3). When A is the stronger ad overall (so 28.1 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 $\widehat{\Delta}^{\text{Targ}}_{\text{AB}}$ reflects a Simpson's reversal?** They won't. The advertiser observes the estimated *aggregate* lifts $\widehat{\lambda}^{\text{Targ}}_{\text{A}}$ and $\widehat{\lambda}^{\text{Targ}}_{\text{B}}$, but none of the true user-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 ($\rho_Y \approx 1$). [12] For example, if the Words and Numbers ads are differentiated by, say, the use of "two" vs "2" in ad copy, but with no other distinguishing characteristics, then their response rates among the audience should be similar. 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

[12] Small effect sizes are consistent with results of large meta-analyses in online advertising (Johnson et al. 2017b), social media advertising (Gordon et al. 2019), and television advertising (Lodish et al. 1995; Shapiro et al. 2020).

then uses that copy as the basis for delivering Words almost solely to Poets and Numbers almost solely to Quants ($\rho_\tau \gg 1$), then the targeting policy 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 the Words ad will be stronger among both types might only target users who were assigned to Words, and none who were assigned to Numbers. But an experiment to compare Words and Numbers needs to expose at least some users to Numbers, even though it is the weaker ad overall. Given the high degree of heterogeneity between types, Quants assigned to Numbers will still outperform Poets assigned to Words, so in an experiment the algorithm might aim to get as many conversions as it can from Numbers out of the Quants. *The requirements of the experimental design could force (or at least lightly nudge) the algorithm toward targeting policies that actually make a Simpson's reversal more likely.*

- **What if there are more than two user types?** Although our analysis uses the simple two-type, two-ad case, concerns about a Simpson's reversal are still relevant when the number of user types is larger. In the general case with n_{X} user types, a Simpson's reversal happens when, $\lambda_{x_1\text{A}} > \lambda_{x_1\text{B}}$, $\lambda_{x_2\text{A}} \geq \lambda_{x_2\text{B}}$, $\ldots$, $\lambda_{x_{n_{\text{X}}}\text{A}} \geq \lambda_{x_{n_{\text{X}}}\text{B}}$, (with strict inequality for at least one type), but $\widehat{\lambda}_{\text{A}}^{\text{Targ}} < \widehat{\lambda}_{\text{B}}^{\text{Targ}}$ in aggregate. While the antecedent conditions for a "pure" Simpsons reversal cannot arise if any of the $\lambda_{x_j\text{B}} > \lambda_{x_j\text{A}}$ 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.

Simulation

Next, we demonstrate through a simulation study how divergent delivery and response heterogeneity conspire to cause the gap between $\widehat{\Delta}_{AB}^{\text{Targ}}$ and Δ_{AB}^{Aud} ($\widehat{\mathscr{E}}_{AB}^{\Delta} \neq 0$). The simulation will reveal how those effects are moderated by conditions described by the sets of ratios $\{\alpha_Y, \pi_Y, \rho_Y\}$ and $\{\alpha_\tau, \pi_\tau, \rho_\tau\}$. We will also show how a Simpson's reversal can occur when the algorithm *overtargets* based on small differences between ads.

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* test with holdout with three ads $Z \in \{A, B, C\}$, to which users in the audience are randomly assigned with equal probabilities ($\zeta_A = \zeta_B = \zeta_C = 1/3$). Additionally, the simulation invokes the following assumptions:

- An outcome is 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 probability of user being targeted with any ad is $\widetilde{\Phi} = \mathbf{P}(\tau = 1) = .2$, and acts as a budget constraint. The conversion probability for all users, across all ads, is $\widetilde{\Theta}^{(1)} = \mathbf{P}\left(Y^{(1)} = 1\right) = .2$.[13]
- The probability of being randomly assigned to the holdout arm, $\mathbf{P}(R = 1)$, is same for all users.

[13] Although the orders of magnitude of targeting and response probabilities are larger than one might see in practice, the same patterns would arise just by using larger simulated audiences.

Table 2: Definitions of Simulated Quantities.

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

The "experimental conditions" of the study are selections from $\rho_{Y} \in \{1, 8\}$ 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 $2n_{Z} = 6$ 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 $\bar{Y}_{Z,\text{Exp}}^{(1)}$ and $\bar{Y}_{Z,\text{Hold}}^{(0)}$ are computed by tallying corresponding potential outcomes of targeted users who are in the respective treatment arm (exposed) and holdout arm arms. Because we have also simulated *all* of the potential outcomes for *all* users, we can also infer both $\bar{Y}_{Z,\text{Aud}}^{(1)}$ and $\bar{Y}_{Z,\text{Aud}}^{(0)}$ for the entire audience, regardless of users' ad assignment or exposure status, including counterfactual potential outcomes. The "true" effects among all simulated users in an audience are λ_{A}^{Aud}, λ_{B}^{Aud}, and Δ_{AB}^{Aud}, while $\widehat{\lambda}_{B}^{\text{Targ}}$, $\widehat{\lambda}_{B}^{\text{Targ}}$, and $\widehat{\Delta}_{AB}^{\text{Targ}}$ mimic the estimated effects that the platform would compute on behalf of the advertiser. Simulated values of $\widehat{\mathscr{E}}_{A}^{\lambda}$, $\widehat{\mathscr{E}}_{B}^{\lambda}$, and $\widehat{\mathscr{E}}_{AB}^{\Delta}$ follow. Table 2 defines effects and biases in terms of these simulated proportions.

Figures 7 and 8 reveal how heterogeneity and targeting policies interact to lead to different signs and magnitudes of biases. The y-axes show $\widehat{\mathscr{E}}_{A}^{\lambda}$ and $\widehat{\mathscr{E}}_{B}^{\lambda}$ in the top and middle rows of panels, and $\widehat{\mathscr{E}}_{AB}^{\Delta}$ in the bottom rows. The nature of the bias is driven by targeting and is moderated by heterogeneity. The logic of any one of these panels is as follows.

- **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}). In the left column ($\rho_{Y} = 1$), there is no user-ad response interaction, so the the ratio of response propensities of Poets to Quants is the same for users assigned to each ad. In the right column ($\rho_{Y} = 8$), the user-ad response interaction generates higher lift propensities for Poets exposed to the Words ad and Quants exposed to

the Numbers ad, than for Poets exposed to Numbers and Quants exposed to Words. For all panels in Figs. 7 and 8, the Words ad A is stronger than Numbers ad B in aggregate across user types ($\alpha_Y > 1$). 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 ($\pi_Y > 4/3$).[14]

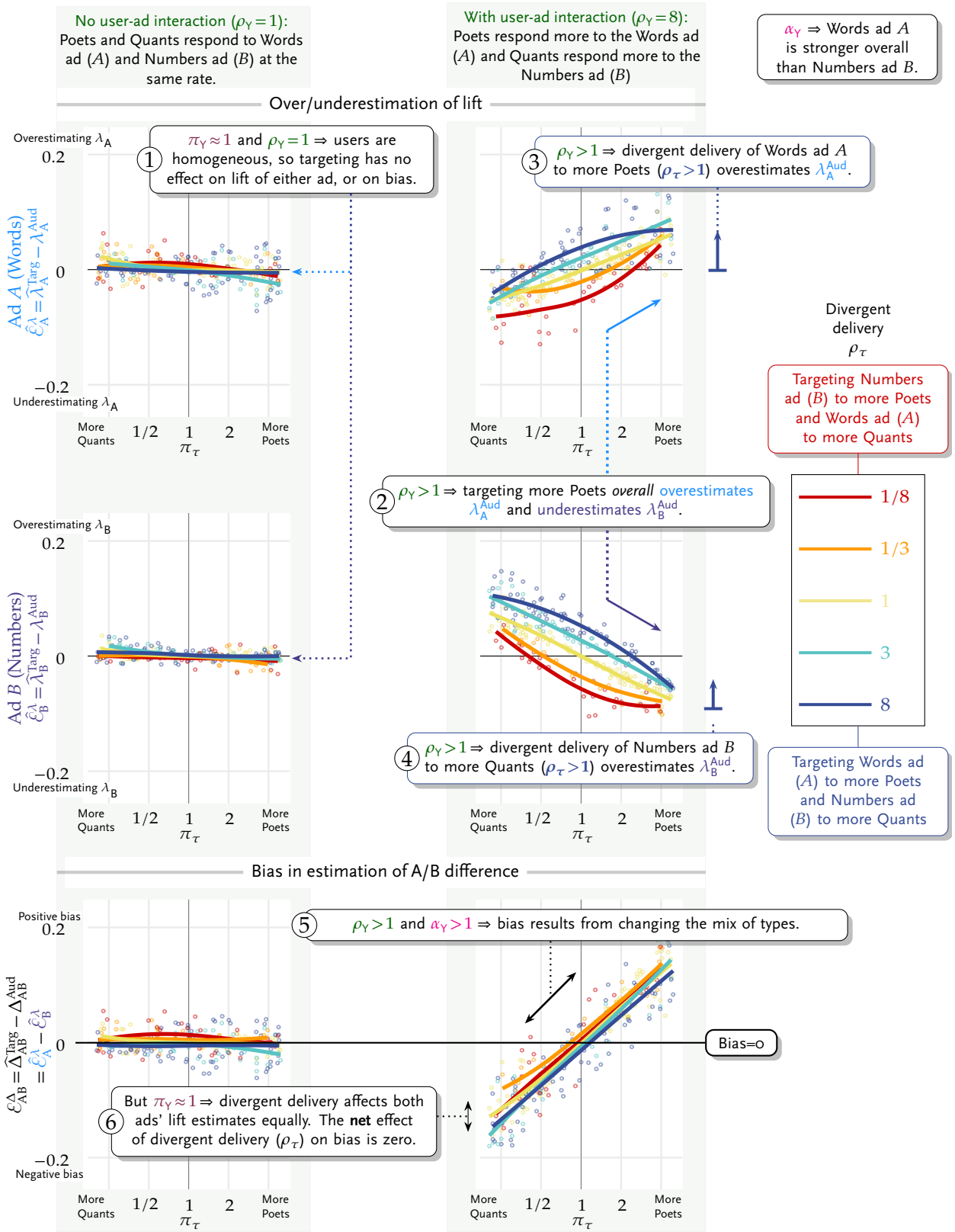
- **Within panels in Figs. 7 and 8, parameters governing targeting decisions differ.** Targeting policies differ for audiences within a panel in two ways. First, type-specific targeting (π_τ, continuous on the x-axis) describes mixes ranging from more Quants ($\pi_\tau < 1$) to more Poets ($\pi_\tau > 1$), 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.

To highlight the simulated audiences with the most extreme effects, and show how different parameter values attenuate those effects, we examine five examples of parameter combinations and their outcomes in Figs. 7 and 8.

Example 1 (Fig. 7, right column; $\alpha_Y > 1$, $\pi_Y \approx 1$, $\rho_Y = 8$). 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 heterogeneous: audiences have extreme user-ad interaction ($\rho_Y = 8$), where Poets exposed to Words and Quants exposed to Numbers have higher lift propensities than their marginal effects α_Y and π_Y would suggest. In this right column, the top panel describes the bias in lift for Words ad A ($\widehat{\mathscr{E}}^{\lambda}_{A}$). Changing the mix of the targeted users through π_τ (x-axis) and ρ_τ (color) affects the gap between $\widehat{\lambda}^{\text{Targ}}_{A}$ and λ^{Aud}_{A}. We start by considering when the algorithm targets more Poets than Quants ($\pi_\tau > 1$, right side of x-axis; e.g., $\pi_\tau = 4$ implies a $4:1$ ratio of Poets to Quants), compared to the audience mix ($\gamma_P = 1/2$ implies a $1:1$ ratio). When the algorithm also

[14] A $\rho_Y = 1/8$ audience where the user-ad response interaction favors Poets exposed to Numbers and Quants exposed to Words would generate symmetric outcomes, so we decided to save space by not including plots for those audiences.

Figure 7: Simulated $\widehat{\mathscr{E}}^{\lambda}_{A}$, $\widehat{\mathscr{E}}^{\lambda}_{B}$, and $\widehat{\mathscr{E}}^{\Delta}_{AB}$ when A is More Effective than B ($\alpha_Y > 1$) and $\Theta^{(1)}_{P} \approx \Theta^{(1)}_{Q}$ ($\pi_Y \approx 1$).



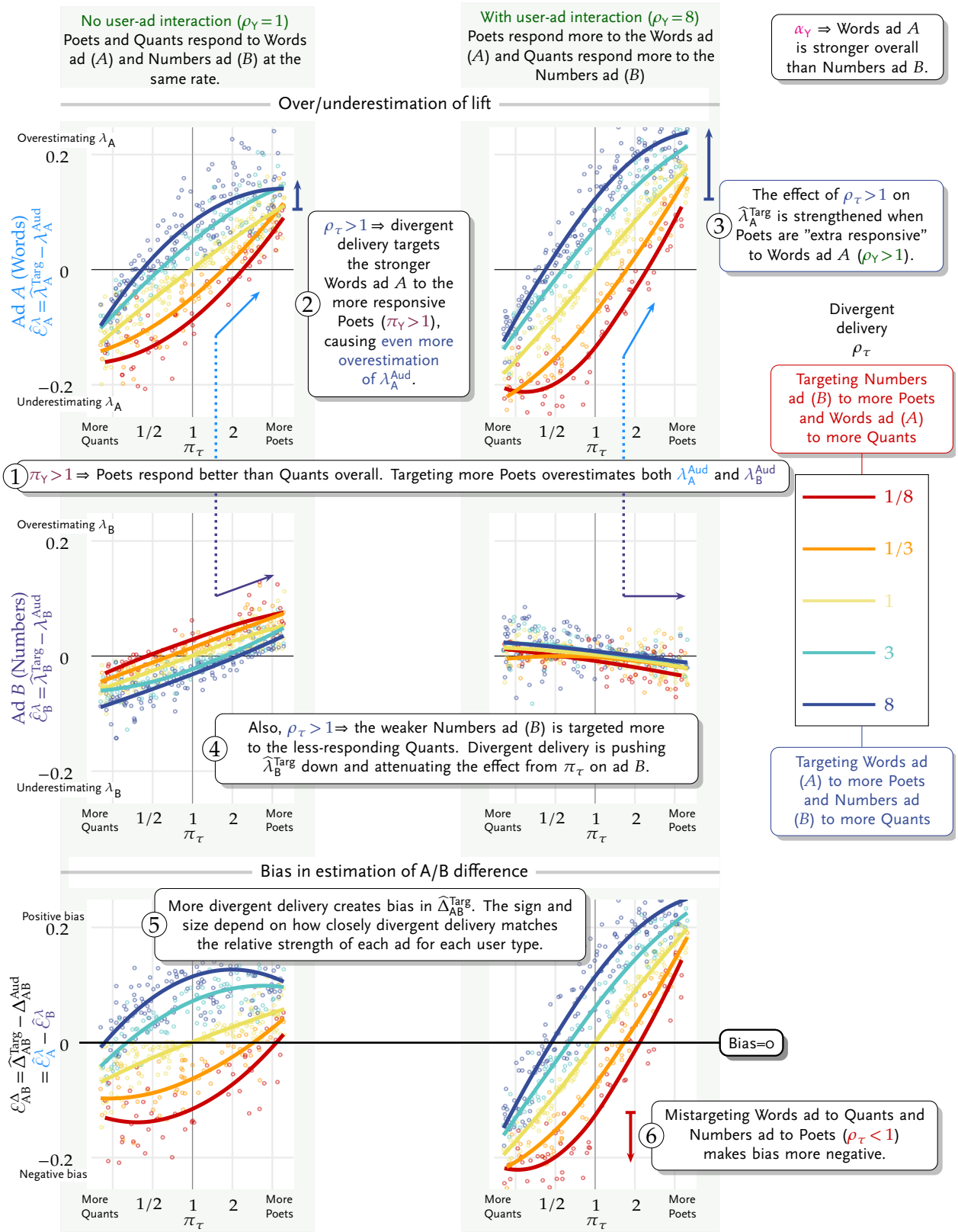
Note: Follow the numbered signposts.

employs divergent delivery favoring A-Poets and B-Quants ($\rho_\tau = 8$; blue line), the resulting mix of users targeted with A skews more to the ad's best responders, A-Poets over A-Quants, than in the mix in the audience. This pushes the estimate of the lift for the targeted mix ($\widehat{\lambda}_{\text{A}}^{\text{Targ}}$) above the true lift for the audience ($\lambda_{\text{A}}^{\text{Aud}}$), so the aggregate lift of ad A is overestimated ($\widehat{\mathscr{E}}_{\text{A}}^{\lambda} > 0$).

For the Numbers ad B, the same response patterns and targeting policies have the opposite effect on bias ($\widehat{\mathscr{E}}_{\text{B}}^{\lambda} < 0$). Unlike in the top-right panel where $\rho_\tau > 1$ and $\pi_\tau > 1$ both favored Words (A), in the middle-right panel this same targeting policy of $\pi_\tau > 1$ (right side of x-axis) and $\rho_\tau = 8$ (blue) creates two opposing effects on Numbers (B). On the margin, targeting (through $\pi_\tau > 1$) favors Poets, who do not respond as well to the Numbers ad. This in turn depresses $\widehat{\lambda}_{\text{B}}^{\text{Targ}}$. But the divergent delivery aspect of the targeting policy (through $\rho_\tau > 1$) favors targeting the better responding Quants who were assigned to Numbers, which increases $\widehat{\lambda}_{\text{B}}^{\text{Targ}}$. As the mix becomes less dominated by Poets ($\pi_\tau > 1$ but decreasing right to left), the lift estimate approaches the audience value ($\widehat{\mathscr{E}}_{\text{B}}^{\lambda} = 0$). Then, as it becomes more dominated by Quants ($\pi_\tau \leq 1$), the targeted mix overestimates the true audience lift ($\widehat{\mathscr{E}}_{\text{B}}^{\lambda} > 0$). As a result, the bias in the estimated difference in the lifts ($\widehat{\mathscr{E}}_{\text{AB}}^{\Delta} = \widehat{\mathscr{E}}_{\text{A}}^{\lambda} - \widehat{\mathscr{E}}_{\text{B}}^{\lambda} = \widehat{\Delta}_{\text{AB}}^{\text{Targ}} - \Delta_{\text{AB}}^{\text{Aud}}$; bottom right 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 Web 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 ($\pi_{\text{Y}} \approx 1$), divergent delivery (ρ_τ) does not affect the bias $\widehat{\mathscr{E}}_{\text{AB}}^{\Delta}$, which is why the colored lines overlap in the bottom row of Fig. 7.

Example 2 (Fig. 7, middle column; $\alpha_Y > 1$, $\pi_Y \approx 1$, $\rho_Y = 1$). The left column of Fig. 7 shows a case where users' responses are entirely homogeneous. While Words (A) is still stronger than Numbers (B) overall ($\alpha_{\text{Y}} > 1$), now user types respond similarly not only in aggregate ($\pi_{\text{Y}} \approx 1$), but also to each ad ($\rho_{\text{Y}} = 1$). Without response heterogeneity from either π_{Y} or ρ_{Y}, all targeting

Figure 8: Simulated $\widehat{\mathscr{E}}^{\lambda}_{A}$, $\widehat{\mathscr{E}}^{\lambda}_{B}$, and $\widehat{\mathscr{E}}^{\Delta}_{AB}$ when A is More Effective than B ($\alpha_Y > 1$) and $\Theta^{(1)}_P > \Theta^{(1)}_Q$ ($\pi_Y > 1$).



Note: Follow the numbered signposts.

decisions (all combinations of α_τ, π_τ and ρ_τ) equally cause no bias in aggregate lifts ($\widehat{\mathscr{E}}^{\lambda}_{\mathrm{A}} \approx \widehat{\mathscr{E}}^{\lambda}_{\mathrm{B}} \approx 0$), and therefore, no bias in A-B difference in lifts ($\widehat{\Delta}^{\mathrm{Targ}}_{\mathrm{AB}} \approx \Delta^{\mathrm{Aud}}_{\mathrm{AB}}, \widehat{\mathscr{E}}^{\Delta}_{\mathrm{AB}} \approx 0$).

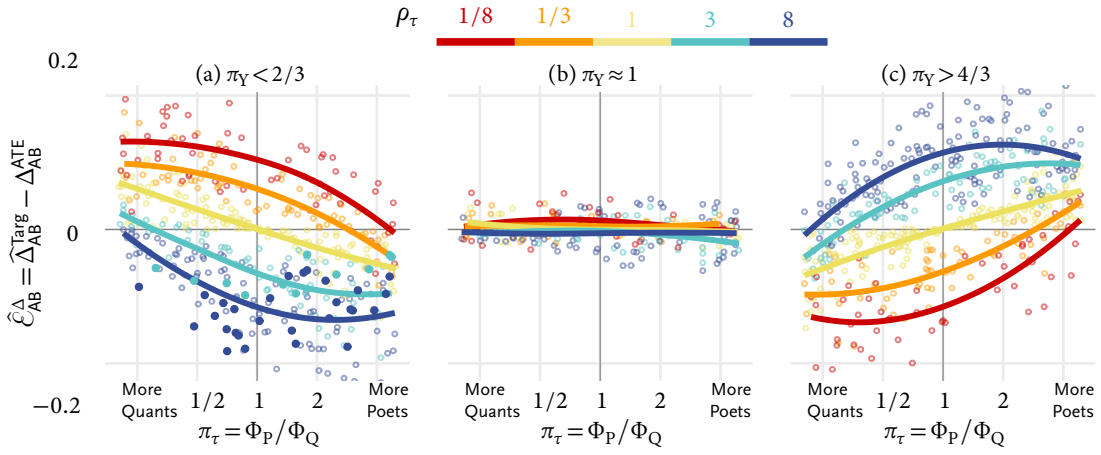
Example 3 (Fig. 8, right column; $\alpha_Y > 1$, $\pi_Y > 1$, $\rho_Y = 8$). For a third example we turn to Fig. 8, whose panels all still have the Words ad A stronger than Numbers ad B on average ($\alpha_{\mathrm{Y}} > 1$). But now there is marginal user heterogeneity where Poets respond better than Quants on average ($\pi_{\mathrm{Y}} > 1$).

The top right panel describes the bias in lift ($\widehat{\mathscr{E}}^{\lambda}_{\mathrm{A}}$) for audiences with high user-ad response interaction ($\rho_{\mathrm{Y}} = 8$), where Poets exposed to Words and Quants exposed to Numbers 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 $\widehat{\lambda}^{\mathrm{Targ}}_{\mathrm{A}}$ and $\lambda^{\mathrm{Aud}}_{\mathrm{A}}$. To see why, consider the most extreme targeting policy shown ($\pi_\tau = 4, \rho_\tau = 8$), where the Words ad ($A$) is delivered more heavily to Poets, and those same A-Poets are exactly the users who have the strongest response to Words. Because 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 the Words ad will be higher among its targeted mix than when averaged across the audience. Therefore, $\widehat{\mathscr{E}}^{\lambda}_{\mathrm{A}} > 0$.

Under the same targeting policy ($\pi_{\mathrm{Y}} = 4$, $\rho_\tau = 8$), the Numbers ad (B) will be delivered to more Quants than Poets, and even more heavily to Quants who were assigned to Numbers. 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 Numbers is weaker than Words overall. But the user-ad response interaction effect in $\rho_{\mathrm{Y}} = 8$ points in the other direction — the Quants who are exposed to the Numbers ad have a greater response rate than the marginal effects alone dictate. Among the Poets exposed to Numbers, the response rate exhibits similar offsetting forces: Poets respond better on average, but have a lower response because of their assigned ad. As a result, the targeting policy creates only minimal bias for ad B ($\widehat{\mathscr{E}}^{\lambda}_{\mathrm{B}}$ is small). Therefore, the bias in estimated difference in lifts, $\widehat{\mathscr{E}}^{\Delta}_{\mathrm{AB}}$ is dominated by $\widehat{\mathscr{E}}^{\lambda}_{\mathrm{A}}$.

In Fig. 9, three panels illustrate the interesting case of audiences with no user-ad response inter-

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



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

actions ($\rho_Y = 1$) even though there may be heterogeneity 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 ($\widehat{\Delta}^{Targ}_{AB} \approx \Delta^{Aud}_{AB}, \widehat{\mathscr{E}}^{\Delta}_{AB} \approx 0$).

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 the Words ad and Quants seeing the Numbers ad ($\rho_\tau = 8$; blue). The Words lift is overestimated as the most responsive users are "doubly targeted" not just because they are Poets, but because they are Poets who were randomly assigned to the Words ad (i.e., a targeting policy with high π_τ and ρ_τ). At the same time, the lift of the Numbers ad (*B*) is only slightly underestimated due to offsetting forces: among users assigned to Numbers, a high ρ_τ means the policy targets more Numbers-liking Quants, but the high π_τ still means that more Poets (who don't respond to Numbers as well) are targeted. 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, left panel; $\alpha_Y > 1$, $\pi_Y < 1$, $\rho_Y = 1$; Simpson's Reversal). The *filled* circles in Fig. 9a identify audiences whose simulated estimates meet the conditions for a Simpson's reversal, where aggregation bias can cause $\widehat{\Delta}_{AB}^{Targ}$ to have a different sign from the Δ_{AB}^{Aud} the advertiser is trying to infer. Compared to Fig. 9c ($\pi_Y > 4/3$), Fig. 9a ($\pi_Y < 2/3$) 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 the Numbers ad much better that Poets (Factor Eq. 28.2 is positive and large). And because $\alpha_Y > 1$ and $\rho_Y = 1$, the Words ad is the stronger ad among both Poets and Quants (Factors Eq. 28.1a and Eq. 28.1b are both positive, but not necessarily large). The remaining 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 Words, lowering the estimate of the aggregate lift of Words relative to its true value in the audience. At the same time, the mix of users targeted with the Numbers ad contains more of the better responding Quants than are in the audience, so the effect of the Numbers ad is overestimated. If this divergence in mixtures is strong enough, then σ_{QA} will be small relative to σ_{QB}, making Eq. 28.3 (and the entire LHS of Eq. 28) large. So as long as the difference in the true ads effects, within user types, is small, the RHS of Eq. 28 will be small enough that the estimated aggregated lift for Numbers will be higher than Words even though both user types respond better to Words than Numbers.

General Insights From the Simulation.

Looking across the bottom rows of Figs. 7 and 8, and Fig. 9, we summarize the most important patterns of bias in the *A*-*B* difference ($\widehat{\mathscr{E}}_{AB}^{\Delta}$).

- More extreme values of bias ($\widehat{\mathscr{E}}_{AB}^{\Delta}$) appear when the user-ad interacton is aligned with the marginal effects. That is, when user response is strongest among Poets ($\pi_Y > 1$) and for ad *A* ($\alpha_Y > 1$) in aggregate, response is especially strong among *A*-Poets and *B*-Quants ($\rho_Y > 1$).
- 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}, and indicates the *incremental bias* generated by divergent delivery ($\widehat{\mathscr{E}}^{\Delta}_{\text{AB}} \mid \pi_\tau, \rho_\tau$), relative to what the bias would have been had divergent delivery been disabled ($\widehat{\mathscr{E}}^{\Delta}_{\text{AB}} \mid \pi_\tau, \rho_\tau = 1$). This effect from divergent delivery is greatest when types are otherwise targeted equally at the margin ($\pi_\tau = 1$). 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 $\sigma_{\text{PA}} \to 1$ and $\sigma_{\text{PB}} \to 1$. Therefore, $\lambda^{\text{Targ}}_{\text{A}} \to \lambda_{\text{PA}}$ and $\lambda^{\text{Targ}}_{\text{B}} \to \lambda_{\text{PB}}$. So while estimates $\widehat{\lambda}^{\text{Targ}}_{\text{A}}$, $\widehat{\lambda}^{\text{Targ}}_{\text{B}}$, and $\widehat{\Delta}^{\text{Targ}}_{\text{AB}}$ 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.

- In the absence of both types of user-level differences ($\pi_{\text{Y}} = 1$ and $\rho_{\text{Y}} = 1$), no amount of targeting will create bias $\widehat{\mathscr{E}}^{\Delta}_{\text{AB}}$. 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 ($\pi_{\text{Y}} = 1$) 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 ($\rho_{\text{Y}} \neq 1$), but does not entirely eliminate the bias caused by marginal user heterogeneity.

Discussion

While the *presence* of algorithmic targeting and its potential implications for online experimentation are not unknown in the marketing research community, this paper is the first to formally show the anticipated *consequences* of divergent delivery on how *advertisers* should interpret *comparisons between ads* from *aggregated results* of so-called "randomized" multi-ad lift studies. Our contribution to the literature goes beyond showing how algorithmic ad targeting causes the mixes of *targeted* users to differ from the audience, and across ads. That is merely an antecedent of the bias. And while researchers like Eckles et al. (2018) and Ali et al. (2019) have documented specific ex-

amples where ads are targeted to users based on user characteristics that are observable to the researchers, advertisers do not have that luxury of detecting those patterns when the targeting criteria are unknown and unobservable to the advertisers themselves.

Advertisers who run *A*-*B* tests online ought to know what they are getting themselves into. This paper is the first to conceptualize the problem from the point of view of the advertiser instead of the platform. Our analysis describes what the nature of bias will actually turn out to be, in terms of both sign and magnitude, under particular targeting policies, patterns of user heterogeneity, and aggregation rules. It uncovers how 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. And instead of just stating that a confound exists, we show the conditions in which it will be a problem, and why. For comparisons between ads, this problem is not resolved by random holdout approaches that are otherwise effective for testing the effectiveness of a single ad. 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.

Perhaps the most striking implication of our paper is the potential for an undetectable Simpson's reversal when platforms aggregate experimental results across undisclosed targeting criteria. 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. While Simpson's Paradox is commonly presented as a statistical curiosity,

online advertising creates realistic conditions for it that have not previously been discussed in the literature.

We have some simple advice for advertisers and researchers who experiment to learn about how their audience responds to creative content, but are considering running these 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 pecuniary objective. The same targeting algorithm that improves lift for a single ad simultaneously distorts estimates of the difference in lift between ads. If the goal of the experiment is to scientifically compare responses to creative content between ads, isolated from the effects of the targeting algorithm, then aggregated results from online experimentation platforms that target users during the experiment will lead to biased inferences.

However, our advice for advertisers whose concerns are to predict which version of an ad will "do best" in a non-experimental campaign under similar conditions on the same targeting ad platform is different: *Carry on*. An ad's overall "performance" depends not only on creative content, but how the targeting algorithm operates on that content to target certain users with that ad. As long as the advertiser's experimental objective does not rely on separating those drivers of performance, the biases we discuss in this paper should cause no problem. Still, those advertisers need to know what effects that *A*-*B* comparison is, and is not, measuring.

Unfortunately, there is nothing the research-oriented advertiser can do by itself to mitigate the bias in *A*-*B* difference estimates caused by algorithmic targeting. All of the power lies with the publisher, which has complete control over the design, specifications, and parameters of experiments it lets researchers run on its platform. 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. That is, it could reduce bias in causal inference by "disabling" the divergent delivery aspect of the targeting algorithm during the course of an experiment. Importantly, disabling divergent delivery does not remove the *A*-*B* comparison bias entirely. In Figs. 7 to 9, that bias lies

along the yellow $\rho_\tau = 1$ lines, which are still non-zero when user types are targeted on the margin ($\pi_\tau \neq 1$), and responses to the campaign are heterogenous (either $\pi_Y \neq 1$ or $\rho_Y \neq 1$). But a smaller proportion of exposed users who are likely to convert restricts the number of short-term revenue-generating opportunities for both the publisher and advertiser. Thus, disabling divergent delivery would effectively increase the cost of running the experiment. 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 during experiments, and advertisers accept this.

Put even more strongly, we doubt the publisher has any incentive to reduce the kind of bias we discuss in this paper at all. One of the reasons publishers provide experimental tools is to show off how well the platform generates value. The publisher wants the *A*-*B* difference among the *targeted* users to be accurate because it makes more money when the advertiser runs effective targeted ads *on that publisher's platform*. Those are the results that can be extrapolated to a non-experimental "production" campaign on that platform, and any throttling of targeting means the algorithm is not seen in the best possible light. Letting the advertiser opt out of targeting during an experiment, or providing advertisers more finely-grained reports of results (essentially converting some the unobservable dimensions of heterogeneity to observable by the advertiser) would not only reveal the true effectiveness of the targeting algorithm, but could also generate privacy concerns. Also, inferences about the effectiveness of creative elements among a broader population can be more generalizable to other media channels than estimates only for targeted users or for a common mix of users. 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. The incentives of the platform and the experimenter are fundamentally misaligned (de Langhe and Puntoni 2021), so advertisers should not expect publishers to offer to run less-biased experiments on the advertiser's behalf anytime soon.

Finally, we note that external validity is imperfect in nearly every marketing research experiment. Whether the subject pool consists of students in an introductory marketing class, participants in a crowdsourced task management marketplace, or users of an online social networking application

with certain predefined characteristics, the representativeness of that subject pool to a greater population of interest is always a question. But as we emphasized, experiments conducted on online targeted advertising platforms are not like other experiments. In our framework, the pre-specified audience *is* the population of interest, and we make no claims about bias when generalizing beyond that audience. But in more traditional lab experiments, the researcher has at least some information about how the characteristics of the subject pool might deviate from the population. The advertiser in a targeted online ad experiment cannot do that, other than to describe the subjects as "whomever the targeting algorithm decided to target." So while online experimental platforms let the advertiser define a reference population, results are applicable to that reference population only to the extent that the set of *targeted* users is representative. And the advertiser has no way of knowing how unrepresentative the targeted users might be.

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.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91(434):444–455.

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.

De Langhe, B. and S. Puntoni (2021). "Does Personalized Advertising Work as Well as Tech Companies Claim?" *Harvard Business Review*. URL: https://hbr.org/2021/12/does-personalized-advertising-work-as-well-as-tech-companies-claim.

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.

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.

Johnson, G. A. (2021). "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.

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

Rosenbaum, P. R. and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70(1):41–55.

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.

Web Appendix

Relationship Between Divergent Delivery and Posterior Mixtures in Figure 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.

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

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. 24,

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

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. 25 and 26 for λ_{PA} and λ_{QA},

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

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

Substituting into Eq. W.2,

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

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

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

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

Equation W.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. W.7).

Simulation Details

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 $\Theta^{(0)}_{PA} \leftarrow \Theta^{(1)}_{PC}$, $\Theta^{(0)}_{QA} \leftarrow \Theta^{(1)}_{QC}$, $\Theta^{(0)}_{PB} \leftarrow \Theta^{(1)}_{PC}$, $\Theta^{(0)}_{QB} \leftarrow \Theta^{(1)}_{QC}$, $\Theta^{(1)}_{PC} \leftarrow \Theta^{(1)}_{PC}$, and $\Theta^{(0)}_{QC} \leftarrow \Theta^{(1)}_{QC}$.
- Sample marginal ratios α_τ, α_Y, π_τ, and π_Y.[15]
  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 $\overline{\pi}_\tau \leftarrow \min\left(\overline{\pi}_\tau,\ \frac{1}{6\widetilde{\Phi} - \Phi_{PC} - \Phi_{QC} - 1}\right)$.

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

8. If $\Theta^{(1)}_{\text{PC}} + \Theta^{(1)}_{\text{QC}} < 6\widetilde{\Theta}^{(1)} - 1$, then adjust $\underline{\pi}_{\text{Y}} \leftarrow \max\left(\underline{\pi}_{\text{Y}},\ 6\widetilde{\Theta}^{(1)} - \Theta^{(1)}_{\text{PC}} - \Theta^{(1)}_{\text{QC}} - 1\right)$ and $\overline{\pi}_{\text{Y}} \leftarrow \min\left(\overline{\pi}_{\text{Y}},\ \frac{1}{6\widetilde{\Theta}^{(1)} - \Theta^{(1)}_{\text{PC}} - \Theta^{(1)}_{\text{QC}} - 1}\right)$.

9. Sample $\pi_\tau \sim \log_2\text{Unif}\left(\underline{\pi}_\tau, \overline{\pi}_\tau\right)$ and $\pi_{\text{Y}} \sim \log_2\text{Unif}\left(\underline{\pi}_{\text{Y}}, \overline{\pi}_{\text{Y}}\right)$

- Solve for the remaining elements of the audience profile.[16]

10. Set the following intermediate values.

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

$$S_Y \leftarrow \sqrt{(\alpha_{\text{Y}}\pi_{\text{Y}} - 1)^2 + (\alpha_{\text{Y}} - \pi_{\text{Y}})^2\rho_{\text{Y}}^2 + 2\rho_{\text{Y}}(\alpha_{\text{Y}}\pi_{\text{Y}}(\alpha_{\text{Y}} + \pi_{\text{Y}} + 4) + \alpha_{\text{Y}} + \pi_{\text{Y}})}$$

$$F_\tau \leftarrow (\alpha_\tau + 1)(\pi_\tau + 1)(\rho_\tau - 1)$$

$$F_Y \leftarrow (\alpha_{\text{Y}} + 1)(\pi_{\text{Y}} + 1)(\rho_{\text{Y}} - 1)$$

11. Set the remaining targeting probabilities.

$$\Phi_{\text{PA}} \leftarrow \frac{2\widetilde{\Phi}}{F_\tau}(\rho_\tau(\alpha_\tau + \pi_\tau + 2\alpha_\tau\pi_\tau) - \alpha_\tau\pi_\tau - S_\tau + 1)$$

$$\Phi_{\text{PB}} \leftarrow \frac{2\widetilde{\Phi}}{F_\tau}(\pi_\tau(\rho_\tau - 2) - \alpha_\tau(\pi_\tau + \rho_\tau) + S_\tau - 1)$$

$$\Phi_{\text{QA}} \leftarrow \frac{2\widetilde{\Phi}}{F_\tau}(\alpha_\tau(\rho_\tau - 2) - \pi_\tau(\alpha_\tau + \rho_\tau) + S_\tau - 1)$$

$$\Phi_{\text{QB}} \leftarrow \frac{2\widetilde{\Phi}}{F_\tau}(\rho_\tau(\alpha_\tau + \pi_\tau + 2) + \alpha_\tau\pi_\tau - S_\tau - 1)$$

12. Set the remaining conversion rates.

$$\Theta^{(1)}_{\text{PA}} \leftarrow \frac{2\widetilde{\Theta}^{(1)}}{F_Y}(\rho_{\text{Y}}(\alpha_{\text{Y}} + \pi_{\text{Y}} + 2\alpha_{\text{Y}}\pi_{\text{Y}}) - \alpha_{\text{Y}}\pi_{\text{Y}} - S_Y + 1)$$

$$\Theta^{(1)}_{\text{PB}} \leftarrow \frac{2\widetilde{\Theta}^{(1)}}{F_Y}(\pi_{\text{Y}}(\rho_{\text{Y}} - 2) - \alpha_{\text{Y}}(\pi_{\text{Y}} + \rho_{\text{Y}}) + S_Y - 1)$$

$$\Theta^{(1)}_{\text{QA}} \leftarrow \frac{2\widetilde{\Theta}^{(1)}}{F_Y}(\alpha_{\text{Y}}(\rho_{\text{Y}} - 2) - \pi_{\text{Y}}(\alpha_{\text{Y}} + \rho_{\text{Y}}) + S_Y - 1)$$

$$\Theta^{(1)}_{\text{QB}} \leftarrow \frac{2\widetilde{\Theta}^{(1)}}{F_Y}(\rho_{\text{Y}}(\alpha_{\text{Y}} + \pi_{\text{Y}} + 2) + \alpha_{\text{Y}}\pi_{\text{Y}} - S_Y - 1)$$

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

- For each audience, simulate user-level data.

  13. Sample $Y_{\mathrm{Z}}^{(1)} \sim \text{Bernoulli}\left(\Theta_{\mathrm{XZ}}^{(1)}\right)$ for all users and all Z, conditional on user type X_i.

  14. Sample $Y_i^{(0)} \sim \text{Bernoulli}\left(\Theta_{\mathrm{X}}^{(0)}\right)$ for all users, conditional on user type X_i.

  15. Sample $\tau \sim \text{Bernoulli}(\Phi_{\mathrm{XZ}})$ for all users, conditional on user type X_i and assigned ad Z_i.

  16. Sample $R \sim \text{Bernoulli}(\mathbf{P}(R = 1))$ 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: $\bar{Y}_{\mathrm{Z,Exp}}^{(1)}$, $\bar{Y}_{\mathrm{Z,Hold}}^{(0)}$, $\widehat{\lambda}_{\mathrm{Z}}^{\mathrm{Targ}}$, and $\widehat{\Delta}_{\mathrm{AB}}^{\mathrm{Targ}}$.

  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: $\bar{Y}_{\mathrm{Z,Aud}}^{(1)}$, $\bar{Y}_{\mathrm{Z,Aud}}^{(0)}$, $\lambda_{\mathrm{Z}}^{\mathrm{Aud}}$, and $\Delta_{\mathrm{AB}}^{\mathrm{Aud}}$.

  19. Compute the bias for the audience: $\widehat{\Delta}_{\mathrm{AB}}^{\mathrm{Targ}} - \Delta_{\mathrm{AB}}^{\mathrm{Aud}}$.

For each of these conditions, we simulated 200 audience profiles using the algorithm above. Using the $\Theta_{\mathrm{XZ}}^{(0)}$, $\Theta_{\mathrm{XZ}}^{(1)}$, 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 Table 2 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.