Where A-B Testing Goes Wrong: How Divergent Delivery Affects What Online Experiments Cannot (and Can) Tell You About How Customers Respond to Advertising

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

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

July 30, 2024

Accepted for publication at *Journal of Marketing*

Abstract

Marketers use online advertising platforms to compare user responses to different ad content. But platforms' experimentation tools deliver different ads to distinct and undetectably optimized mixes of users that vary across ads, even during the test. Because exposure to ads in the test is non-random, the estimated comparisons confound the effect of the ad content with the effect of algorithmic targeting. This means experimenters may not be learning what they think they are learning from ad A-B tests. The authors document these "divergent delivery" patterns during an online experiment for the first time. They explain how algorithmic targeting, user heterogeneity, and data aggregation conspire to confound the magnitude, and even the sign, of ad A-B test results. Analytically, the paper extends the potential outcomes model of causal inference to treat random assignment of ads and user exposure to ads as separate experimental design elements. Managerially, the authors explain why platforms lack incentives to allow experimenters to untangle the effects of ad content from proprietary algorithmic selection of users when running A-B tests. Given that experimenters have diverse reasons for comparing user responses to ads, the authors offer tailored prescriptive guidance to experimenters based on their specific goals.

Keywords: Targeted online advertising, experimental design, A-B testing, measuring advertising effectiveness, Simpson's paradox, causal inference, field experiments, algorithmic targeting, social media platforms

---

The authors gratefully acknowledge comments and guidance on this and earlier versions of this paper from Eric Bradlow, Bart De Langhe, Dean Eckles, Brett Hollenbeck, Lacey Jeffrey, Garrett Johnson, Joshua Lewis, Blake McShane, Stefano Puntoni, Nils Wernerfelt, and many seminar and conference participants. The authors also recognize the contributions of Hye Jin Yoon to the development of the experimental design and advertising copy used in the empirical example in this paper. The SMU University Research Council provided financial support for this research.

Many online advertising platforms (e.g., Meta, Google) allow marketers to conduct A-B tests to learn how users respond when exposed to different ads. The platforms provide A-B testing tools that ostensibly randomize certain aspects of ad delivery, where some users are exposed to ad A and others to ad B. But these tools are not randomizing exposure to ads in a way that allows the experimenter to learn the causal effects of ad content on user response, isolated from the effect of the targeting algorithm itself. This is because targeting algorithms serve each ad to different mixes of users optimized for each ad, even during the course of the experiment.

We call this pattern *divergent delivery* because the mixes of types of users targeted with each ad diverge from each other.[1] A consequence of divergent delivery is that the A-B comparison estimated from the data reflects the combination of effects from both ad content and algorithmic selection of users, which is different than what would have occurred under random exposure. That means experimenters may not be learning what they think they are learning from the A-B tests of their ads. Whether this state of affairs is a problem depends on the reasons for running the experiment.

This paper is a conceptual introduction, reference guide, and tutorial to the issues surrounding divergent delivery in online advertising experiments. First, we give background on targeted ads and online experiments, and then move onto practical realities of running ad A-B tests. Next, we provide an empirical illustration of divergent delivery using an A-B test conducted in the field. Then we formalize a framework for describing targeting policies, user responses, and the relationships among them, and illustrate how *algorithmic targeting, user heterogeneity, and data aggregation all conspire to confound the magnitude, and even the sign, of ad* A-B *test results*. Finally, we provide guidance to experimenters on consequences of divergent delivery, and consider when divergent delivery does or does not matter for various objectives.

Background On Targeted Advertising and Online Experiments

To understand the implications on experimental results from non-random exposure generated by A-B testing tools, which platforms let experimenters use free of charge beyond the cost of delivering the ads, we first need to understand the more general context of targeted advertising. In essence, an

[1] Johnson (2023) introduced the term *divergent delivery*. Ali et al. (2019) use *skewed delivery* to describe the same patterns.

ad is a bundle of *creative elements*, such as message, copy, and images, that constitute the *content* of the ad. The platform considers delivering ads to members of an *audience* of users, which is the population of users that the advertiser specifies along demographic dimensions provided by the platform. *Exposure* refers to the platform's successful presentation of the ad on the user's screen, regardless of the user's behavior. A subset of that audience will be exposed to the ad.

Relevance is a standard term of art that describes the combination of the platform's expectations of user response, preferences, and behavior when exposed to an ad, as well as its overall assessment of *ad quality*. Relevance is determined at the user-ad level, and it is the elemental driver of divergent delivery, where different ads are targeted to different mixes of types of users.

Determining which users in an audience are exposed to an ad is the crux of online ad *targeting*. The operation of user-level ad targeting relies on relevance because delivering more relevant ads to users enhances the overall user experience (i.e., reduces irrelevant ads that could drive users away from the platform), and makes it more likely that users will click, like, convert, or buy after exposure to an ad (i.e., resulting in greater revenue for the advertiser and platform).

In providing a targeted advertising service, the platform offers a bundled two-part value proposition to advertisers. First, the platform is a vehicle to expose ad impressions to users, "selling eyeballs," much like any other advertising channel. Second, the platform also offers use of a targeting algorithm to place the "right" ads in front of the "right" users. Targeting provides value to the advertiser because the initial audience can be quite large, and budgets are limited. It is neither feasible nor cost efficient for ads to be exposed to the entire broadly audience, and it may also not be desirable to the advertiser for the exposed users to be uniformly random subset of the broadly defined audience.

Instead, algorithms allocate ads to specific users through what can be understood conceptually as a *targeting process*. Under the hood, the "engine" of the targeting process is typically an auction, where advertisers place bids for the right to show ads to users in an audience. But the winner of an auction for the right to place an ad on a particular user's screen is not based on monetary value of the bids alone, but also the ad content and user-ad relevance.[2] The precise inputs and methods that

[2]For instance, see Google's documentation defining Auction (https://bit.ly/GAaucDef) and Ad Quality Score (https://bit.ly/GAaqDef). For Meta's definition, see https://www.bit.ly/MetaAboutAdDelivery.

determine the relevance of ads to users, how relevance influences auction results, and thus, which users are targeted with each ad, are proprietary to particular platforms and are not observable to advertisers.[3] For this paper, regardless of the specifics of how any ad platform implements these ideas, what matters is the effect of these processes: that particular ads are targeted to particular users, using information about users and ads not reported to advertisers.

This application of relevance to the targeting process is inextricably embedded in all online advertising platforms. Targeting does not take place without considering relevance. In fact, the impact of user-ad relevance on divergent delivery cannot merely be disabled in certain contexts, even during A-B tests.

Once a targeted user is exposed to an ad, the advertiser is hoping for a beneficial *outcome*, like a click, page view, or conversion. The platform reports *results* for each ad that are aggregated outcomes across users, typically up to the level of coarsely defined demographic groups, like age or gender. Note that the advertiser does not observe outcomes at the user level.

Single-Ad Illustration

To illustrate, consider a landscaper, whose designs focus on native plants and water conservation. For now, suppose the landscaper advertises through an online platform with just one ad. The creative elements in this ad highlight the aesthetic aspects of this style of outdoor design, showing decorative rocks, water features, and little grass. We will refer to this as the "aesthetics ad". The landscaper's desired outcome is to generate quality leads suitable for follow-up by a sales team, and considers its focal market segment to be high-income homeowners over 30 years old within a 50 mile radius of Houston. When configuring the online ad campaign, this market segment constitutes the audience of the ad campaign.[4]

As with heterogeneous customers in a market segment, users who are members of the same audience may have different propensities for generating an outcome after being exposed to an ad. So how does the platform assign different relevance scores to different users eligible to be exposed to an ad?

[3] See https://patents.google.com/patent/US10325291B2 for one publicly disclosed approach.

[4] In the language of experimental design, the "audience" is the population of interest from which experimental subjects are selected. The marketer's analogy is the "target segment" for their strategy. We avoid that phrase because "target" has a different meaning in the context of online ad delivery, and use "market segment" instead.

We characterize this heterogeneity by allowing the platform to implicitly describe each user with an unobservable latent *type*. A user's type is determined by all of the information collected by the platform, and is used to influence which ads are exposed to which users. The type of any particular user is understood by the algorithm, but is unobservable to the advertiser because it is based on the platform's proprietary information. In this paper, type is an abstraction for how a platform might distinguish users who may find an ad to be highly relevant from users who are much less likely to engage with the ad at all.

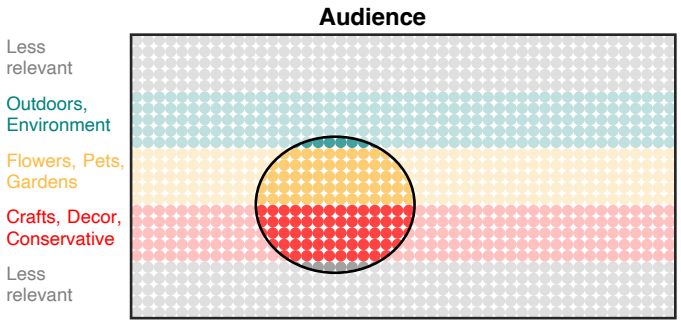
Targeting helps maximize the effectiveness of the landscaper's limited ad budget by exposing the ad to those users most likely to engage with the ad. For example, suppose that among our landscaper's defined audience, the platform algorithm identifies a type of user who comments on photos of nature, has friends who support environmental causes, and has searched for adventure-style outdoor vacations. This type of user may be more likely to be interested in the kinds of services this landscaper offers. The platform may also identify another type of user who posts to craft and decor affinity groups and frequently visits personal finance websites. Those users may not find the landscaper's aesthetic ad to be relevant at all. The targeting of some types of users more than others comes from both the economic forces in the underlying auction (Lambrecht and Tucker 2019), and how the platform uses ad relevance to modify advertisers' bids to determine auction winners. But how the algorithm determines relevance for types of users is not precisely known, and may not even be able to be enumerated or reproduced by the platform itself (Gordon et al. 2019).

We illustrate the targeting of ads based on user type in Figure 1, where each circle is a user, the collection of users in the rectangle constitutes the audience, bright circles are users who are exposed to the ad (users with dim circles are unexposed), and colors indicate groups of users with similar types[5]

Unbeknownst to the landscaper (but not necessarily against her wishes), the algorithm is targeting users based on user types. These targeted users fall within the "targeting ovals" in Figure 1. In this example, users are exposed if and only if they are targeted, so circles inside the oval are bright and

[5]For expositional purposes, we combined similar user types into color groups to represent heterogeneity within the audience that is unobservable to the advertiser. We display these as discrete levels of one dimension, with the understanding that there may be further heterogeneity within these groups.

Figure 1: Single-ad blob



NOTE: Each colored circle is a user in the specified audience. Colors, which vary vertically, indicate one of five user types, as described in the left margin. The oval's vertical position and the mix of colored circles inside the oval reflect the mix of users. Bright circles are exposed users, and dim circles are unexposed users. Users targeted with their assigned ad are contained inside the respective "targeting oval." Here, all targeted users are exposed (inside the oval, bright), while all untargeted users are unexposed (outside the oval, dim)

those outside are dim (we will relax this condition later). The positioning of the oval shows the mix of user types among the targeted users. For this example, the algorithm is implicitly predicting that an ad highlighting the aesthetic aspects of this landscaper's design style would be more relevant to users interested in flowers, general gardening, crafts and decor (red and yellow types) than those interested in outdoors and the environment (green type), or others who are even less interested in the landscaper's services (gray types). The audience includes both users who were expected to engage with the ad, and users who are not interested in gardening at all. But there is still heterogeneity among our discrete groups, and the landscaper's ad budget constrains the number and mix of users who can be targeted. So the targeted users are not just red and yellow users; some are green or gray.

As a result of this targeting process, the targeted users are not *representative* of the focal market segment. We see this in Figure 1 where the distribution of colors inside the targeting oval is different from the distribution in the full rectangle. The landscaper may wonder if an observed increase in sales leads means that an ad strategy focusing on landscape aesthetics is effective for their entire the audience (focal market segment), or only the subset of the audience that the algorithm targeted, based on auction results and the relevance of the ad to those users. Beyond the definition of the audience, the landscaper cannot describe precisely how the mix of user types in the audience differs from the mix among the targeted users because the factors that influence whether a user is targeted with this ad are not observable.

A-B tests and Divergent Delivery

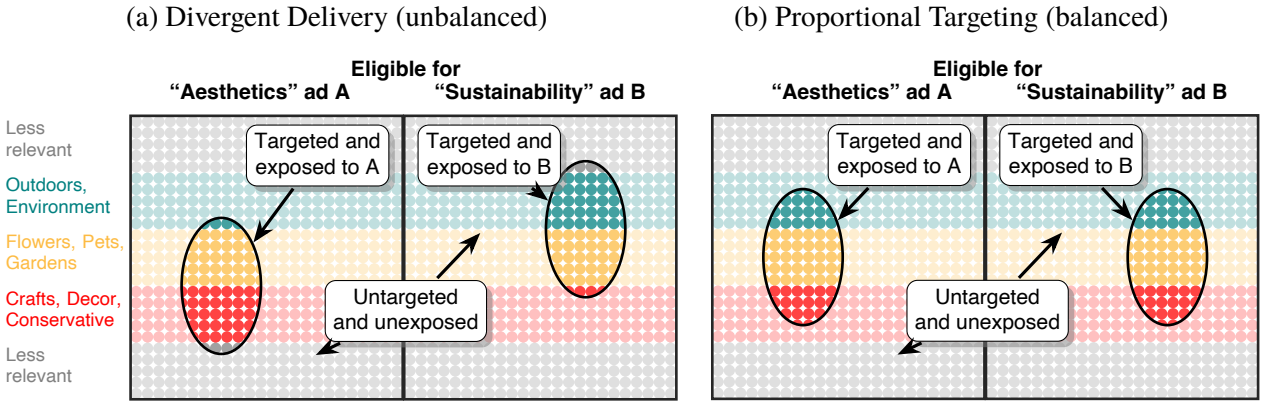
Now that we have introduced the concepts of relevance, targeting, and user types in the context of a single ad, we can consider an experiment that compares the effects of multiple ads. An A-B test is a specially configured form of ad delivery where the advertiser is now an *experimenter* trying to learn how exposure to *different ads* stimulates user outcomes. Just as in a single ad campaign, targeted users are not *representative* of the audience because users are targeted based on the relevance of ad contact to different user types. But with an A-B test, the types of users exposed to different ad treatments are not *balanced*. That is, not only are the targeted mix and audience mix different, but the targeted mix for one ad differs from the target mix for the other ad. This is a critical concern when interpreting the results of an A-B test conducted in a targeted ad environment.

Continuing with the example from above, our landscaper wants to gain insights into the brand positioning of their business. For example, suppose it wants to know if it should it present these services as an approach to landscaping that is attractive and decorative, or one that is more sustainable and environmentally friendly? To help set its marketing strategy, it conducts an A-B test where ad A is the same aesthetics-focused ad from above. Ad B is a different ad that highlights how landscaping with native plants that consume less water is more sustainable (the "sustainability ad"). The results of this test will influence all aspects of the marketing mix, including images and copy for online and offline marketing, choice of advertising channels, and refinement of the focal market segment.

We illustrate two possible designs for this A-B test in Figure 2, one under the real-world conditions of divergent delivery (users targeted with different ads have different type mixes), and one under a hypothetical *"proportional targeting"* policy (where type mixes are balanced across ads). In both panels, the rectangle represents the same audience as in Figure 1. But in an A-B test, audience users are first randomly assigned to be *eligible* to be exposed to either A or B. This random assignment is shown by the partitioning of the audience into two sides. Because this eligibility is random, the mix of colors on either side of the partition is the same for both ads. Users who are eligible for each ad and are subsequently targeted with that ad are shown inside their respective ads' targeting ovals.[6]

[6]We assume the total number of targeted users is the same for both ads, so the total area of the A and B ovals equals the area of the targeting oval in Figure 1.

Figure 2: A-B Tests with and without Divergent Delivery



NOTE: Users are randomly assigned to be eligible to see either ad A or B (left/right rectangles within each panel). The mix targeted with each ad differs across ads in (a) Divergent Delivery, but these mixes are the same across ads in (b) Proportional Targeting.

In Figure 2a, the aesthetics ad A is targeted to the same mix of user types as in Figure 1. But suppose the algorithm predicts that a different mix of user types will be more likely to engage with the sustainability ad B. For instance, ad B is expected to be more relevant to liberal-leaning, outdoorsy-types of users (green) because B highlights how this landscaper's practices are more environmentally sustainable. The algorithm may also infer that the politically conservative users interested in home decor (red users) are less likely to respond to B, so very few of those users are in the B targeting oval. Thus, the targeting oval for users eligible to view B is positioned to cover a different mix of user-type colors than the oval for A. Being targeted with an ad is a precondition of being exposed to that ad, so the distribution of types among users exposed to an ad will *diverge* between ads A and B.

Divergent delivery is an inevitable consequence of targeting ads to users based (in part) on relevance of the content of the ads to users, because some users will find ad A more relevant than B, and others will find B more relevant than A. Targeting with divergent delivery has two key implications for the mix of targeted users. First, targeted users are *unrepresentative* of the advertiser-chosen audience: users targeted with the aesthetics ad A are not representative of the audience eligible for ad A, and users targeted with the sustainability ad B are also not representative of audience eligible for ad B. Second, the mix of users targeted with A differs from the mix of users targeted with B. That means the two mixes of users targeted with the aesthetics and sustainability ads are *unbalanced*.

To understand the distinction between *representativeness* and *balance*, contrast the divergent delivery design in Figure 2a with the proportional targeting design in Figure 2b. In a proportional targeting design types of users would be targeted and exposed to both ads in the same proportions (the mixes of colors inside the targeting ovals are the same for both ads). Targeted users are still not representative of the audience because the algorithm is incorporating relevance when determining which users see which ads. But there is *balance* in user types exposed to the experimental treatments. This is the kind of design that might result if a platform were to "disable" divergent delivery during an A-B test, and is analogous to a randomized control trial among targeted users.

Proportional targeting, however, is not likely to appear in practice. By constraining the mix of users targeted with the aesthetic ad and the sustainability ad to be the same, the platform loses a degree of freedom. To illustrate, the mix of users targeted with the aesthetic ad is different between Figure 2b, when it is constrained to be equal to another ad's mix, and Figure 1, when it is unconstrained. To the extent that divergent delivery targets the users most likely to engage with that particular ad, proportional targeting would lead both the platform and the advertiser to leave money on the table during the course of the A-B test (as we demonstrate later). Whether the experimenter can accept that tradeoff depends on the reasons for running the experiment.

Interpreting A-B Test Results Under Divergent Delivery

The representativeness and balance of an A-B test determine how the results of the test should be interpreted, and what the experimenter can learn about ad content. Consider the hypothetical proportional targeting design (with no divergent delivery) in Figure 2b. The lack of representativeness among those targeted users means that the experimenter cannot extrapolate inferences about the effect of ad content to the entire audience. Insights that the landscaper might have wanted to draw from the experimental results (i.e., whether to position the brand around aesthetics or sustainability) would not be generalizable to the entire focal market segment.

But among those targeted users in Figure 2b, the mixes of user types are balanced. Like a randomized control trial, the observed difference in outcomes between users exposed to the two ads is attributable only to differences in the content of the ads. This balance means that the experimenter can interpret

the A-B comparison to be a *causal* effect of the different creative elements of the ads. The following caveats apply, which limit the use of any causal inference from this data: the lack of representativeness means results of this randomized A-B test are applicable only to the types of users who were selected for the experiment (i.e., targeted), and the criteria used to select those targeted users are unobservable to the experimenter (de Langhe and Puntoni 2021; Braun et al. 2024).

In practice, divergent delivery in targeting makes causal inference about the effect of ad content impossible because the comparison in outcomes between ads is not "apples to apples" (different colors in the ovals in Figure 2a). This imbalance occurs because the algorithm considers relevance of ad content when targeting users. Without balance, the ad effects are confounded by the selection of users seeing those ads (Hardisty and Weber 2020; D.'Angelo and Valsesia 2023).

For example, the users exposed to the sustainability ad may consist of a greater proportion of users who comment on nature-related posts (and users similar to them), while users seeing the aesthetics ad will contain a greater proportion of users who post photos of arts and crafts (and users similar to them). The critical issues are: (1) arts and crafts enthusiasts and nature-commenters may be affected by exposure to the creative content of the aesthetics and sustainability ads differently; and (2) the aggregated results of the A-B test, as reported to the experimenter, are not broken down by these granular user types that are applied in the targeting process but are unobservable to the experimenter. The lack of balance might prompt the experimenter to ask: Are users who saw the sustainability ad responding because of their reaction to water-efficient sustainable gardening practices portrayed in the ad? Or is it because the nature-focused individuals who were targeted with the sustainability ad are more likely to react positively to any of the landscaper's ads (because of their baseline interest in the brand) than the other users who were targeted with the aesthetics ad? That is, might the targeted sustainability-focused users have a higher baseline preference for the company and a higher responsiveness to its ads than the targeted aesthetics-focused targeted users have for the company and its ads?

But there is no way for the experimenter to separate how much of the reported A-B difference is due to differences in ad content from how much is due to having different types of users see each ad. The methods for targeting ads to user types are proprietary, and the criteria that drive how the targeted sets

of users for each ad differ from one another and from the intended population are not disclosed to the experimenter. The criteria may not even be saved by the platform for subsequent analysis (Gordon et al. 2019, p. 220). Because A-B test results are aggregated over these unobserved user types, the experimenter is blind to the full extent to which non-random exposure is happening, and there is no way to account for divergent delivery in analyses of the results provided by A-B testing tools.

Holdout Tests

In common discourse and in the literature, the term A-B test is sometimes used for a different kind of experiment: a *holdout test*. The goal of a holdout test is to estimate an ad's *lift*, the incremental value generated by users' exposures to an ad relative to not being exposed, which could then be used as an input to computing the return-on-investment (ROI) of an ad campaign.[7]

An early strategy in online advertising for estimating an ad's lift was to compare its outcome to a "placebo control," such as a public service announcement (PSA) (Lewis et al. 2011; Barajas et al. 2016; Johnson et al. 2017a). The idea was that exposure to a PSA that is unrelated to the focal ad should lead to the same outcomes as not being exposed to any ad at all. But to the targeting algorithm, the PSA is just another ad to be targeted to certain users. The mix of users targeted with the PSA may be different than the mix targeted with the focal ad. Therefore, comparisons between the PSA and the focal ad carry the same warnings about divergent delivery as the A-B comparisons we have discussed so far. The focal ad will be relevant and delivered to some types of users, and the PSA would be relevant and delivered to other types. The resulting lack of balance of user types between the ads would confound the estimate of the focal ad's lift itself.

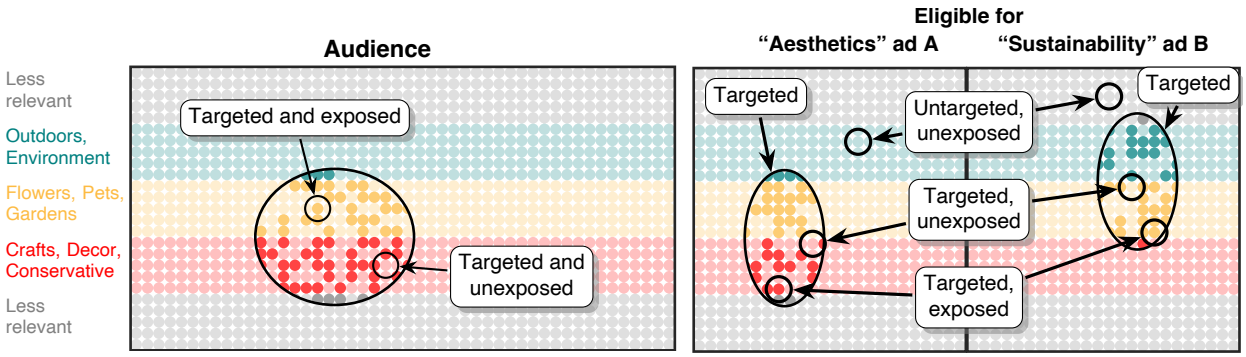
Johnson et al. (2017a), Eckles et al. (2018), and Gordon et al. (2019) explain the merits of estimating lift with a *holdout test* on the focal ad, where targeted users are randomly assigned to one of two *arms* of the experiment: (1) a *treatment arm* whose users are targeted with and exposed to their assigned ad (bright circles inside the oval); and (2) a *holdout arm* whose users are targeted with, but unexposed to, their assigned ad (dimmed circles inside the targeting oval). Instead of receiving their

[7]We follow Ascarza (2018) by defining lift as an incremental difference in ad response. We acknowledge that some researchers define lift in percentage terms (Gordon et al. 2019).

Figure 3: Holdout (lift) tests

(a) Single-ad Holdout Test

(b) A-B test with Holdout



NOTE: As in Figures 1 and 2, users targeted with their assigned ad are contained inside the respective "targeting ovals." But now, a fraction of targeted users (here, 50%) are randomly "held out" and remain unexposed (inside the ovals, dim). Inside each targeting oval, bright and dim circles have the same mixes of colors, indicating balance between targeted exposed and targeted unexposed users.

assigned ad, users in the holdout arm are exposed to a "shadow control" ad (or a "ghost ad" as in Johnson et al. 2017a). The shadow control is the ad the user would have seen had the experiment not been running (typically, the second place ad in the auction).[8] The shadow control may be different for every user, as it represents a "what if" condition of non-exposure as a point of comparison to the treatment ad. Although the treatment and holdout arms are sometimes described as experimental conditions A and B, our use of A and B always refer to different ads. The shadow control ad is a more appropriate baseline than a placebo PSA for measuring lift because it is the true counterfactual condition: which ad would the holdout users have seen if the focal treatment ad did not exist?

Figure 3a illustrates a holdout test on a single ad. Random assignment among the targeted users ensures a single-ad holdout test is balanced because the mixes of user types the same in both arms (the bright and dim circles inside the targeting ovals have the same mix of colors). Therefore, the estimated lift of an ad from holdout tests has a causal interpretation among those targeted users.

An A-B *test with holdout* involves running holdout tests for ads A and B in parallel and comparing the ad-specific lifts to one another. Figure 3b illustrates an example of an A-B test with holdout that is conducted under divergent delivery. The mixes of users in the audience, and among targeted

[8]Different platforms have their own names and implementations for similar A-B testing concepts, like a Conversion Lift experiment at Google: `https://bit.ly/GoogleConvLift`. Meta previously named this a "Conversion Lift Test," but renamed it to "A/B Test with Holdout." We use the term "shadow control" because it is generic and platform-agnostic.

users, are the same as in Figure 2a, but now there is another step to the exposure process. Data are collected from four groups of users who are (1) targeted with and exposed to A; (2) targeted with but not exposed to A; (3) targeted with and exposed to B; and (4) targeted with but not exposed to B. An experimenter would choose this design to compare the *incremental* effects of two ads. Platforms that implement this design are essentially running "two-armed mini randomized experiments" among only the users who were targeted with each ad (Johnson et al. 2017a; Gordon et al. 2019).[9]

One application of an A-B test with holdout is to compare the ROI of two or more ads. However, while the estimation of lift or ROI for each ad *separately* comes from a balanced experiment, the comparison *between* ads does not. This is because the lifts for the two ads are computed from different mixes of user types. So while the calculated ROI for ad A is a causal estimate for users targeted with ad A, and while the calculated ROI for ad B is a causal estimate for users targeted with ad B, the A-B comparison between ROI of A and ROI of B cannot be causally attributed to the difference in the ads' content alone. To summarize, *the confounds caused by divergent delivery across ads are not solved by using a holdout test within each ad.*

Empirical Evidence of Divergence Delivery in Field Experiments On Ad Platforms

Since words like "experiment" and "randomized" appear in A-B testing tools' own documentation, experimenters might reasonably expect the required conditions for causal inference, like balance across experimenter-designed treatments, to also hold for online A-B tests.[10]

But they don't.

Marketers should not be surprised that targeting to specific users based on ad content occurs in online ad campaigns. After all, targeting is part of the platform's basic value proposition to advertisers. What may be surprising to marketers is that targeting with divergent delivery occurs *during* supposedly randomized A-B tests as well. The ubiquity of divergent delivery (Figure 2a) in online advertising means that *experimenters using platforms'* A-B *testing tools cannot assume that exposure to ad treatments is either representative of their chosen audience or balanced across ad treatment groups.*

[9] This design is analogous to a design where the targeting process picks users who are "intended to be treated" (ITT).
[10] www.bit.ly/MetaAboutExperiments

Our interest in this problem was motivated by our real-world experience with the online experiments we conducted in the summer of 2018 while collaborating with the City of Detroit about marketing strategy for employee recruitment. Like many cities in the U.S., the City of Detroit faces challenges in attracting qualified applicants for public service jobs. Searches for candidates use tactics familiar to any marketer: offline ads, events, and social media. The recruiters understood that there are many different reasons that people would want to work for the City, and wanted to learn if different types of messages would encourage individuals to apply who may not be reached by traditional recruitment efforts. The goal was to guide development of advertising content across multiple advertising channels, both online (e.g., search, display ads on multiple websites) and offline (e.g., print, flyers, posters at job fairs). Testing ideas for recruitment campaigns was far easier and cheaper to do via these online ad platforms. Also, city officials wanted to broaden their outreach to candidates that were representative of their residents they serve, while also recognizing the need to concentrate resources on the most likely prospects. As such, we ran these experiments to extrapolate insights to broader marketing decisions, much like the landscaper in our earlier example.

We used Facebook's A-B test with Holdout tool (then known as a "Multi-cell Conversion Lift Test") to test incremental responses to ads with different creative elements (e.g., messaging, images, and copy). The audience for the test consisted of all Facebook and Instagram users, aged 18 to 40, within 20 miles of Detroit City Hall, excluding Canada.[11] Only a small subset of users in this audience were actually targeted with the ads during the experiment, and among those, a holdout group remained unexposed. We constructed a recruiting landing page in collaboration with the City, where the experimental outcome was submitting contact information through an interest form on the webpage. That way the platform could track outcomes for targeted users based on their visits to and actions on our webpage, regardless of whether or not the targeted user was actually exposed to that ad. This enables comparison to the holdout groups.

We tested 14 ads, six of which are shown in Figure 4, with the remaining in Web Appendix A. The ad treatments were created as a $3 \times 2 \times 2$ full-factorial design, plus two placebo control conditions.

[11] We did not restrict the audience based on factors like gender, income, or topical interests because the City wanted to recruit from a population that properly represented residents of Detroit and surrounding communities.

The first factor is the featured career path: data analyst, domestic violence unit officer, and patrol officer. This factor is operationalized in the ad images and the content of the ad copy. The second factor, Other/Self, is reflected in the bolded headlines: "Serve your community and keep Detroit safe" (Other) or "Advance your career and develop your skills" (Self). The third factor, Rational/Emotional appeals, appears in the copy text: "these crimes get solved" (more rational) or "feels safer in her home" (more emotional). The two placebo control ads are PSAs promoting local Detroit non-profit organizations (the Police Athletic League and the Every School Day Counts anti-truancy program) that are unrelated to city employment, but still relevant to the defined audience.

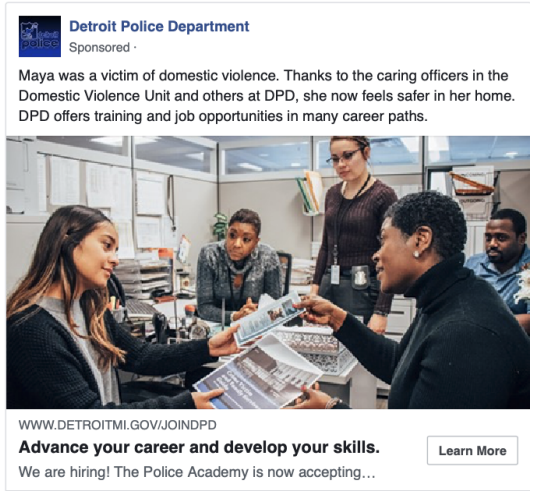
For the purpose of this paper, this study is a demonstration that divergent delivery occurs, at a minimum, along observable dimensions. We include this example as evidence of the lack of balance across ads, so we focus our analysis on patterns in which ad impressions were targeted to which users. During a three-week period, 533,161 impressions were served to 96,150 unique users, 50.2% of whom were female. Figure 5 shows the female proportion of users targeted with each ad. Table WA1 in Web Appendix A breaks down these counts by ad treatment.

Figure 5 reveals a pattern of divergent delivery. That imbalance even reinforces traditional gender roles and stereotypes. First, there is a significant association between gender and career path ($\chi^2_2 = 137$, $p < .001$). The ads featuring the domestic violence officer (with mostly women in the image) were more likely to be targeted to women (52.2% female), while the ads with a male patrol officer in the foreground were targeted more to men (47.4% female). The gender mix of users targeted with the ad featuring a female data analyst was more balanced (49.7% female). Targeted users were also more likely to be females receiving the Emotional appeal than the Rational appeal (50.3% vs 49.2%; $\chi^2_1 = 11.2$, $p = .001$), and the Other-focused ad (50.1%) than the Self-focused ad (49.4%), although the weight of the statistical evidence for that comparison is somewhat weaker ($\chi^2_1 = 3.98$, $p = .046$). We even see a gender disparity between the control ads, where women make up a greater proportion of users targeted with the Every School Day Counts anti-truancy PSA featuring children in a classroom (57.0%) than the PSA for the Police Athletic League featuring showing children on a sports field with male coaches (50.8%; $\chi^2_1 = 42.9$, $p < .001$). Figure 5 also shows some three-way interaction

Figure 4: Six of the 14 Ads from the Detroit Field Experiment

Data Analyst, Self, and Emotional

Domestic Violence Unit, Self, and Emotional



Patrol Officer, Other, and Rational

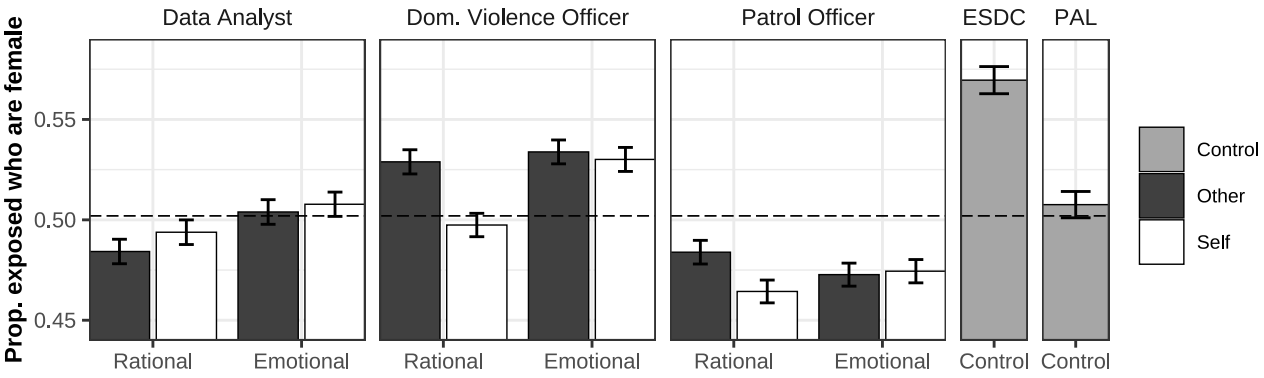
Data Analyst, Other, and Rational

PSA for Every School Day Counts.

PSA for Police Athletic League.

NOTE: Ads come from a full-factorial design with 3(jobs)×2(Self- v. Other-focused)×2(Emotional v.Rational)+2(PSAs). The remaining ads are in Web Appendix A. The authors gratefully acknowledge the contributions of Hye Jin Yoon to the development of the experimental design and advertising copy used in this study.

Figure 5: Observed Imbalance in the Detroit Recruiting Experiment



NOTE: The y-axis is the percentage of females among unique users exposed to an ad (50.2% in aggregate). Ads are grouped in the left three panels by job type. The Rational/Emotional factor (copy text) is on the x-axis, and the Other/Self factor (tag line) is on the color scale. The right two panels with gray bars are control ads: Every School Day Counts (ESDC), and Police Athletic League (PAL). (Error bars = $\pm$ 1 SE).

effects, such as between the Self- and Other-focused ads for users eligible for the domestic violence officer career path with the Rational appeal. Thus, divergent delivery is occurring *even during the course of an experiment*. This had not been documented in the literature prior to this paper.

Detecting this gender imbalance was possible only because the platform released results that were disaggregated by this particular variable. If gender were the only user characteristic the algorithm applied to the targeting decision, we might have been able to correct the imbalance caused by divergent delivery using standard statistical re-weighting techniques. But the targeting algorithms' determinations of relevance depend more on unobservable and proprietary information than on coarse-grained demographics. Even if user types were balanced along demographics, those users would still be unbalanced on the more important latent, unobserved characteristics. Since these results are not broken down by those unobserved factors, there is no way to isolate the effect of the ad content from how the targeting algorithm acts on that content. Thus, even *observable* covariate balance is not sufficient evidence that an A-B test replicates the random assignment of users to ads.

Our firsthand experience with running an A-B test with limited visibility into the data-generating process, including our inability to get useful results from it, motivated us to explore the interplay between targeting and user heterogeneity more deeply in this research. Experimenters like us are mere consumers of a testing platform's A-B testing tool, and they are flying blind as to what might

be driving their results. This is a contrast with an employee of an ad platform who is engaged with the inner workings of their targeting algorithm's infrastructure. The model and analysis that follow are meant to provide more insight into the challenges that these experimenters face when they have no control over how subjects are selected for a study, or how the study participants are assigned to various treatment or control conditions.

Mathematical Framework Characterizing Targeting Policies and User Responses in *A-B* tests

In this section, we more formally describe the interplay of targeting, responses to ad content, and user heterogeneity, and show how divergent delivery affects what experimenters with different objectives can learn from A-B tests in a targeted advertising environment. We will begin by using probabilities to represent the targeting and exposure process, which selects users to be exposed to different ads (experimental treatments). Then, we review the established *potential outcomes* model of causal inference (Rubin 1974), and extend it to accommodate multiple ad treatments. Integrating these two concepts reveals how divergent delivery leads to aggregating experimental outcomes across different probability weights for users exposed to different ads. This framework sets up the analysis in the subsequent section, where we will demonstrate how divergent delivery affects estimates of A-B comparisons for a population of interest, and discuss whether deviations from various baseline estimates are helpful or harmful for experimenters in different marketing objectives.

Targeting Policies and User Types

A targeting policy affects which users are more likely to be selected for exposure to treatments. A key input to a targeting policy is information about the user. Let $X_i = X$, be a categorical variable that defines the latent *type* of user i. This latent type is unobservable to the experimenter but used by the platform's targeting algorithm.[12] For example, one possible value of X might be a composite of several characteristics and behaviors, such as, "male users who receive political memes from friends, post to Instagram at least weekly, and are predicted to purchase gardening supplies in the

[12]We assume that all outcomes and effects are conditional on both membership in the audience and any demographic variables that define aggregation groups in A-B test results (e.g., gender, age, or location). Although these factors influence targeting and outcomes, they are homogeneous within the scope of reported results, so we exclude them from X.

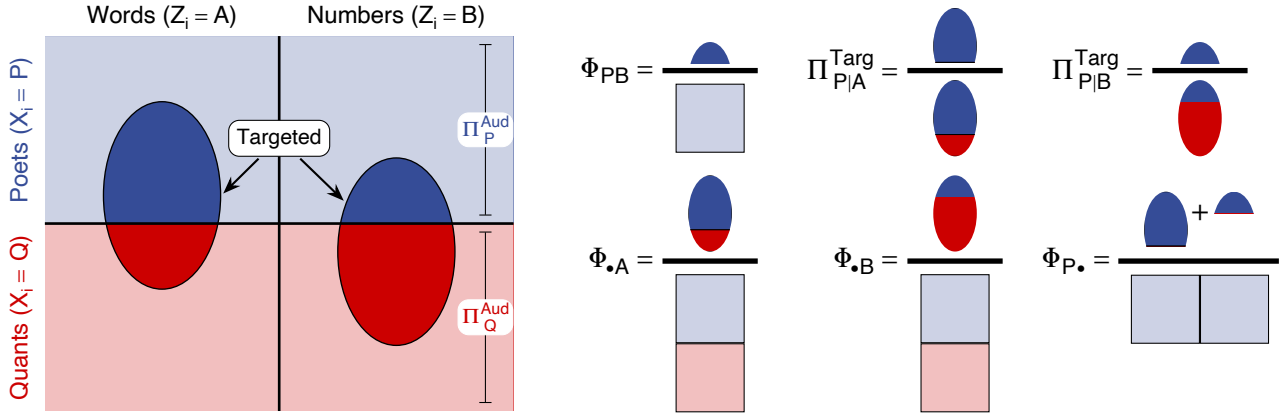
next month." These factors are part of the rich trove of proprietary information about a user that guide the algorithm in determining which creative elements of ads are most relevant to users of that type, and thus which users are most likely to be targeted with which ads. Our use of the user type label X is conceptual, and the entire framework is generalizable to an arbitrary number of types. For simpler exposition, we will often refer to examples with two types of users, P and Q. To make those examples easier to follow, we refer to the $X_i = P$ users as Poets, and the $X_i = Q$ users as Quants.

At the start of an experiment, the platform randomly assigns every user i in the audience to be *eligible* to receive one ad, $Z_i = Z$, and only that ad. We use Z to denote a general ad, and A and B to denote specific ads. As an mnemonic aide, we will often describe A as the "Words" ad, and B as a "Numbers" ad. These conceptual labels let us characterize targeting policies where, for example, more Poets are targeted with the Words ad and more Quants are targeted with the Numbers ad. The proportion of users eligible for each ad is a parameter of the test and does not depend on user type.

After this random assignment process determines eligibility, the algorithm uses the internal information at its disposal to decide whether a user who is eligible to receive ad Z_i will actually be *targeted* with ad Z_i. Some types of users eligible for each ad are more likely to be targeted than others. Because the internal operations of the targeting algorithm are complex, proprietary, and unobservable to the experimenter (as if it were inside a black box), we treat the targeting algorithm *as if* it were a random process, conditional on X_i and Z_i. The implicit probabilities in this model of the targeting process characterize a *targeting policy*, and are set by the algorithm.

The targeting policy itself can be viewed in terms of *targeting probabilities* and *type distributions*, which we define pictorially in Figure 6 and more formally in Table 1. The conditional targeting probability Φ_{XZ} represents how likely it is that the algorithm will target a user of type X who is eligible for ad Z. For example, Φ_{PA} is the proportion of Poets eligible to receive the Words ad (A) who are then targeted. This probability depends on the both X and Z because the algorithm may find the content of one ad relevant for some types of users, but not others. That is, we model targeting

Figure 6: Illustrated Definitions of Targeting Policy Symbols



NOTE: Areas are proportional to user counts, and probabilities are shown as fractions of areas. Targeted users are contained in the darkened ovals. Marginal targeting probabilities $\Phi_{\bullet A}$ and $\Phi_{\bullet B}$ are proportions of columns that are within their respective ovals. Conditional targeting probabilities Φ_{PA} and Φ_{QA} are respective proportions of the left-blue and left-red quadrants that are darkened. Posterior type distributions $\Pi^{Targ}_{P|A}$ and $\Pi^{Targ}_{P|B}$ are proportions of their respective targeting ovals that are blue (for Poets). $\Pi^{Targ}_{P|A} > \Pi^{Targ}_{P|B}$ is evidence of divergent delivery with $\rho_\tau > 1$ (Equation 1).

to be probabilistic at the ad-by-type level.[13] The marginal targeting probabilities $\Phi_{\bullet Z}$ and $\Phi_{X\bullet}$, respectively, refer to targeting probabilities for an ad (aggregated over user types in the audience) and for a type (aggregated over ads).

A *type distribution* is a probability distribution that describes the proportions of a user type among some subset of users. Π^{Aud}_{X} is the proportion of the audience with type $X_i = X$ and, equivalently, the *prior* probability that a randomly chosen user in the audience has that type. We say it is the prior because it exists *before* users are targeted. $\Pi^{Targ}_{X|Z}$ is the type distribution among users who are *targeted* with ad Z and, equivalently, is a *posterior* probability conditional on being targeted with a particular ad. For example, $\Pi^{Targ}_{P|A}$ is the proportion of Poets among users who were targeted with Words (A).[14] When we compare the values of $\Pi^{Targ}_{X|Z}$ (e.g, proportion of Poets among users who were targeted with A compared to proportion of Poets among those targeted with B), we can observe the degree to which the targeted mix for each ad diverges.

[13] Different platforms may implement their experimental platforms in different ways (e.g., the ordering of targeting, eligibility, and exposure processes). The platform's actual targeting algorithm may be deterministic or stochastic, but from the perspective of the experimenter, we can still consider the targeting algorithm *as if* it were random.

[14] By Bayes' Theorem, the relationship between the targeting probabilities and type distributions is $\Pi^{Targ}_{X|Z} = \Pi^{Aud}_{X} \frac{\Phi_{XZ}}{\Phi_{\bullet Z}}$. Targeting is uniformly random across all ads (representative and balanced) if and only if $\Pi^{Targ}_{X|Z} = \Pi^{Aud}_{X}$ for all Z.

Table 1: Symbol Definitions for Targeting Policies and Type Distributions

	Definition		Explanation
	Targeting probabilities (inputs to targeting policy)		
Φ_{XZ}	$= \mathbf{P}\left(\tau_Z^i = 1 \mid X_i = X,\, Z_i = Z\right)$		Probability a type X user who is eligible to receive Z is targeted.
$\Phi_{\bullet Z}$	$= \mathbf{P}\left(\tau_Z^i = 1 \mid Z_i = Z\right)$	$= \sum_{\forall X} \Phi_{XZ} \Pi_X^{Aud}$	Probability a user who is eligible to receive Z is targeted (aggregated over types).
$\Phi_{X\bullet}$	$= \mathbf{P}\left(\tau' = 1 \mid X_i = X\right)$	$= \sum_{\forall Z} \Phi_{XZ} \mathbf{P}(Z_i = Z)$	Probability that a type X user is targeted with whichever ad that user is eligible to receive.
	Type distributions		
Π_X^{Aud}	$= \mathbf{P}(X_i = X)$		Prior probability a user in the audience has type X
$\Pi_{X\|Z}^{Targ}$	$= \mathbf{P}\left(X_i = X \mid \tau_Z^i = 1, Z_i = Z\right)$	$= \frac{\Phi_{XZ}}{\Phi_{\bullet Z}} \Pi_X^{Aud}$	Posterior probability that a user eligible for and targeted with ad Z has type X.
$\Pi_{X\|Z}^{NoTarg}$	$= \mathbf{P}\left(X_i = X \mid \tau_Z^i = 0, Z_i = Z\right)$	$= \frac{1 - \Phi_{XZ}}{1 - \Phi_{\bullet Z}} \Pi_X^{Aud}$	Posterior probability that a user eligible for but *not* targeted with ad Z has type X.
Π_X^{Prop}	Special case of $\Pi_{X\|Z}^{Targ}$ where the type distribution among targeted users is the same for all Z (i.e., under a proportional targeting policy with $\rho_\tau = 1$).		

NOTE: In Tables 1 and 2, $\tau_Z^i = 1$ indicates that user i will be targeted with Z_i when that user is eligible to receive Z_i. If $\tau_Z^i = 0$, then the user cannot be exposed to Z, even if initially eligible to receive it. $\mathbf{P}(Z_i = Z)$ is the proportion of users in the audience who are eligible for ad Z. For brevity, the expression "targeted with ad Z" refers to users who are both targeted with ($\tau_Z^i = 1$) and eligible to see ($Z_i = Z$) ad Z.

Divergent delivery occurs when the proportions of users with different types diverge across users targeted with different ads. We formally define this as the two-way interaction between ad content and user type that results from targeting types of users based on ad content, and equivalently, a posterior odds ratio:

$$\rho_\tau = \frac{\Phi_{PA}}{\Phi_{QA}} \Big/ \frac{\Phi_{PB}}{\Phi_{QB}} = \frac{\Pi_{P|A}^{Targ}}{\Pi_{Q|A}^{Targ}} \Big/ \frac{\Pi_{P|B}^{Targ}}{\Pi_{Q|B}^{Targ}} = \frac{\text{Odds a user targeted with Words is a Poet}}{\text{Odds a user targeted with Numbers is a Poet}} \tag{1}$$

Under a divergent delivery targeting policy with $\rho_\tau > 1$, the extent to which Poets are more likely to be targeted than Quants is greater among users eligible for Words than for Numbers. While the algorithm may choose to target Poets more than Quants overall, and separately may also choose to deliver more Words ads than Numbers ads, this targeting policy means that Poets who are eligible for Words, and Quants who are eligible for Numbers, are more likely to be targeted than whatever their marginal targeting probabilities in isolation would suggest. As a result, Poets are more prevalent among users targeted with Words than among users targeted with Numbers: $\Pi_{P|A}^{Targ} > \Pi_{P|B}^{Targ}$ (visualized

in Figure 6).[15] When $\rho_\tau = 1$, the targeting policy is proportional with no divergent delivery. The posterior type distribution under proportional targeting is $\Pi_X^{Prop} = \Pi_{X|A}^{Targ} = \Pi_{X|B}^{Targ}$, indicating that the mix of targeted users who are Poets vs Quants is balanced across ads.

An Extended Potential Outcomes Model

Following the Rubin (1974) potential outcomes model, a user is an experimental subject who can be in one of two *exposure states*, exposed (1) or unexposed (0). Every user is endowed with a pair of latent *potential* outcomes that represent what the user's outcome would have been in each exposure state: $Y_i^{(1)}$ if exposed, and $Y_i^{(0)}$ if unexposed. These potential outcomes are random variables, where the distributions for $Y_i^{(1)}$ and $Y_i^{(0)}$ may depend on type, but not on the eventual exposure state, of that user. Because the user can be either exposed or unexposed to the ad (exactly one, not both), the user's *realized* outcome Y_i^* is equal to *either* $Y_i^{(1)}$ or $Y_i^{(0)}$. If the user is exposed, then $Y_i^* = Y_i^{(1)}$ is realized, and the other potential outcome, $Y_i^{(0)}$, is a counterfactual: what that user's realized outcome would have been if not exposed. And vice-versa: if a user is unexposed, then $Y_i^* = Y_i^{(0)}$ is realized, and $Y_i^{(1)}$ is the counterfactual. Because the user's exposure state is the only factor determining which of their potential outcomes is realized, $Y_i^{(1)} - Y_i^{(0)}$ is the incremental outcome that is *caused* by exposure to the ad, relative to non-exposure. Most causal inference can be reduced to these kinds of hypothetical "what if" comparisons, which can be captured by differences between potential outcomes.[16]

But the usual way in which the potential outcomes framework has been applied to online ad experiments does not fully capture all of the feasible potential outcomes in an A-B test. For instance, applications like Gordon et al. (2019) are concerned with random assignment to exposure or non-exposure conditions, which is conceptually different from random assignment to the treatments for which a user is eligible to receive (e.g., ads A or B). As a result, $Y_i^{(0)}$ ambiguously could represent a potential outcome for either non-exposure to any ad or exposure to a placebo ad.

To address this limitation, we extend the basic potential outcomes model to capture both dimensions

[15] "Targeting" and "divergent delivery" are different concepts. Targeting makes the type distribution among targeted users unrepresentative of the audience ($\Pi_P^{Prop} \neq \Pi_P^{Aud}$). Divergent delivery is a particular targeting pattern that generates imbalance between users targeted with different ads ($\Pi_{P|A}^{Targ} \neq \Pi_{P|B}^{Targ}$).

[16] From Cunningham (2021, p. 125): "…, potential outcomes is more or less the lingua franca for thinking about and expressing causal statements, and we probably owe D. Rubin (1974) for that as much as anyone."

of exposure and treatment assignments simultaneously and independently. Instead of having only one pair of potential outcomes corresponding to exposure or non-exposure to a treatment, the user is endowed with a $Y^{(1)}_{i,Z}$ and $Y^{(0)}_{i,Z}$ pair for each one of the treatment ads in the experiment. So, for a test with treatment ads A and B, user i has 4 potential outcomes — reflecting 2 (treatments: {A,B}) × 2 (exposure states: {1,0}) — $Y^{(1)}_{i,A}$, $Y^{(0)}_{i,A}$, $Y^{(1)}_{i,B}$, and $Y^{(0)}_{i,B}$. For example, $Y^{(1)}_{i,A}$ is the potential outcome associated with the eligibility for and exposure to ad A, and $Y^{(0)}_{i,B}$ is the potential outcome for when that same user is eligible for ad B, but is not exposed to it. Which of a user's potential outcomes is actually realized as Y^{*}_{i} now depends on two distinct processes: eligibility to receive an ad, and exposure to that ad. While the targeting process determines which potential outcome is ultimately realized by each user, the potential outcomes themselves are latent characteristics of the user do not change.

As in the standard model, causal effects of exposure to an ad is the difference in potential outcomes, but now, these effects can be defined for two ads separately: $Y^{(1)}_{i,A} - Y^{(0)}_{i,A}$ and $Y^{(1)}_{i,B} - Y^{(0)}_{i,B}$. That means we can define the focal causal effect as the difference between the effects of the two ads:

$$\Delta^{i}_{AB} = \left(Y^{(1)}_{i,A} - Y^{(0)}_{i,A}\right) - \left(Y^{(1)}_{i,B} - Y^{(0)}_{i,B}\right) \tag{2}$$

Denoting all potential outcomes related to both ad assignment and exposure state highlights the conceptual distinction between $Y^{(0)}_{i,A}$ and $Y^{(1)}_{i,B}$.

Using this extended framework we can now recognize the importance of the assumption, $Y^{(0)}_{i,A} = Y^{(0)}_{i,B}$, *within a single user* i. We can make this assumption because eligibility for an ad affects behavior only when the given user is exposed to that ad and because behavior does not depend on the assigned ad when the given user is unexposed. Under this assumption, Equation 2 reduces to $Y^{(1)}_{i,A} - Y^{(1)}_{i,B}$, but only at the *individual level for any given user* i. We will return to this point shortly, since considering these quantities across different mixes of users may yield unequal average quantities.

How Targeting, Divergent Delivery, and Heterogeneity Affect *A-B* Test Results

Aggregating Quantities Across Mixtures of Users

The practical implications of divergent delivery for the experimenter comes down to how outcomes are *aggregated across user types*. The basic building blocks for aggregating potential outcomes

Table 2: Expected Potential Outcomes and Treatment Effects

Definition		Explanation
Expected potential outcomes		**Expected response among users …**
$\mu^{(D)}_{XZ} = \mathbf{E}\left[Y^{(D)}_{i,Z} \mid X_i = X \right]$		…with type X and eligible for ad Z.
$\mu^{(D)}_{X} = \mathbf{E}\left[\mathbf{E}\left[Y^{(D)}_{i,Z} \mid Z_i = Z \right] \mid X_i = X \right]$	$= \sum_{\forall Z} \mu^{(D)}_{XZ} \mathbf{P}(Z_i = Z)$	…with type X.
$\mu^{(D)}_{Z,Targ} = \mathbf{E}\left[Y^{(D)}_{i,Z} \mid \tau^i_Z = 1, Z_i = Z \right]$	$= \sum_{\forall X} \mu^{(D)}_{XZ} \Pi^{Targ}_{X\mid Z}$	…eligible for and targeted with ad Z.
$\mu^{(D)}_{Z,NoTarg} = \mathbf{E}\left[Y^{(D)}_{i,Z} \mid \tau^i_Z = 0, Z_i = Z \right]$	$= \sum_{\forall X} \mu^{(D)}_{XZ} \Pi^{NoTarg}_{X\mid Z}$	…eligible for but not targeted with ad Z.
$\mu^{(D)}_{Z,Aud} = \mathbf{E}\left[Y^{(D)}_{i,Z} \right]$	$= \sum_{\forall X} \mu^{(D)}_{XZ} \Pi^{Aud}_{X}$	…in the audience.
Expected differences in potential outcomes		**Lift of ad Z among users …**
$\lambda_{XZ} = \mathbf{E}\left[Y^{(1)}_{i,Z} - Y^{(0)}_{i,Z} \mid X_i = X \right]$	$= \mu^{(1)}_{XZ} - \mu^{(0)}_{XZ}$	…with type X.
$\lambda^{Aud}_{Z} = \mathbf{E}\left[Y^{(1)}_{i,Z} - Y^{(0)}_{i,Z} \right]$	$= \sum_{\forall X} \lambda_{XZ} \Pi^{Aud}_{X}$	…in the audience.
$\lambda^{Targ}_{Z} = \mathbf{E}\left[Y^{(1)}_{i,Z} - Y^{(0)}_{i,Z} \mid \tau^i_Z = 1, Z_i = Z \right]$	$= \sum_{\forall X} \lambda_{XZ} \Pi^{Targ}_{X\mid Z}$	…who are eligible for and targeted with ad Z.
Expected differences-in-differences in potential outcomes		**Expected *A*-*B* difference for users …**
$\Delta^{X}_{AB} = \mathbf{E}\left[\left(Y^{(1)}_{i,A} - Y^{(0)}_{i,A} \right) - \left(Y^{(1)}_{i,B} - Y^{(0)}_{i,B} \right) \mid X_i = X \right]$	$= \lambda_{XA} - \lambda_{XB}$	…with type X.
$\Delta^{Aud}_{AB} = \mathbf{E}\left[\left(Y^{(1)}_{i,A} - Y^{(0)}_{i,A} \right) - \left(Y^{(1)}_{i,B} - Y^{(0)}_{i,B} \right) \right]$	$= \sum_{\forall X} \Delta^{X}_{AB} \Pi^{Aud}_{X}$ $= \lambda^{Aud}_{A} - \lambda^{Aud}_{B}$	…in the audience.
$\Delta^{Prop}_{AB} = \mathbf{E}\left[\left(Y^{(1)}_{i,A} - Y^{(0)}_{i,A} \right) - \left(Y^{(1)}_{i,B} - Y^{(0)}_{i,B} \right) \mid \tau^i_Z = 1, Z_i = Z \right]$	$= \sum_{\forall X} \Delta^{X}_{AB} \Pi^{Prop}_{X}$	…targeted under a proportional targeting policy: $\Pi^{Prop}_{X} = \Pi^{Targ}_{X\mid A} = \Pi^{Targ}_{X\mid B}$

NOTE: See note to Table 1 for definition of eligibility and targeting (τ^i_Z). The expected values of potential outcomes, $Y^{(D)}_{i,Z}$, are defined for $D \in \{0,1\}$.

are $\mu^{(1)}_{XZ}$ and $\mu^{(0)}_{XZ}$, which are *conditional expected values* of respective potential outcomes $Y^{(1)}_{i,Z}$ and $Y^{(0)}_{i,Z}$ among users with type X. Similarly, the *lift*, λ_{XZ}, is the conditional expected causal effect of ad Z among users with type X. The complete sets of expected outcomes and lifts for each level of aggregation and conditioning are formally defined in Table 2.

Any targeted subset on an audience contains a mix of different types of users. Aggregation across user types involves taking weighted mixtures of over the appropriate distributions of user types. There are three such distributions: the mix of the entire audience (Π^{Aud}_{X}), the mix among targeted users ($\Pi^{Targ}_{X|Z}$), and the mix of untargeted users ($\Pi^{NoTarg}_{X|Z}$). For example, among all users targeted with ad A, $\mu^{(1)}_{A,Targ}$ and λ^{Targ}_{A} are weighted averages of $\mu^{(1)}_{XA}$ and λ^{Targ}_{A}, using $\Pi^{Targ}_{X|A}$ as the weights.

Some, but not all, aggregate potential outcomes can be estimated from the experimental results. Let

Table 3: Definitions of Estimates from Observed Data

	Definition	Explanation
$\widehat{\lambda}_{Z}^{Targ}$	$= \bar{Y}^{(1)}_{Z,Targ} - \bar{Y}^{(0)}_{Z,Targ}$	Estimates λ_{Z}^{Targ} if and only if exposed and unexposed users have the same type mix (as in a holdout test).
$\widehat{\lambda}_{Z}^{Conf}$	$= \bar{Y}^{(1)}_{Z,Targ} - \bar{Y}^{(0)}_{Z,NoTarg}$	Confounded estimate of λ_{Z}^{Targ} when unexposed outcomes are estimated from untargeted users.
$\widehat{\Delta}_{AB}^{Targ}$	$= \widehat{\lambda}_{A}^{Targ} - \widehat{\lambda}_{B}^{Targ}$	Difference in estimated lift among targeted users eligible for their respective ads.

$\bar{Y}^{(1)}_{A,Targ}$ and $\bar{Y}^{(1)}_{B,Targ}$ be the estimates of $\mu^{(1)}_{A,Targ}$ and $\mu^{(1)}_{B,Targ}$ that are computed as the respective averages of $Y^{*}_{i} = Y^{(1)}_{i,A}$ and $Y^{*}_{i} = Y^{(1)}_{i,B}$ among users who were targeted with A and B. When the A-B test includes a holdout set, the platform will also provide $\bar{Y}^{(0)}_{A,Targ}$ and $\bar{Y}^{(0)}_{B,Targ}$ for targeted unexposed users whose behavior is tracked by the platform. But outcomes that are observable to the experimenter are only a few of the outcomes that are realized by the user. If the platform does not track outcomes for untargeted users, experimenters will not be able to compute the corresponding $\bar{Y}^{(0)}_{A,NoTarg}$ and $\bar{Y}^{(0)}_{B,NoTarg}$, even though untargeted users realize $Y^{(0)}_{i,A}$ and $Y^{(0)}_{i,B}$. Further, $\bar{Y}^{(1)}_{A,NoTarg}$ and $\bar{Y}^{(1)}_{B,NoTarg}$ are never available in practice because untargeted users are never exposed.

Table 3 defines the estimates of lifts in terms of these observed averages. In a holdout test, the estimated lift among users targeted with A is $\widehat{\lambda}_{A}^{Targ} = \bar{Y}^{(1)}_{A,Targ} - \bar{Y}^{(0)}_{A,Targ}$. It is an estimated *causal* effect as long as the type distributions are the same for targeted exposed users and targeted unexposed users. In Figure 7, this estimated lift is shown as the vertical distance between (A1) and (U1), because targeted users are randomly assigned to the treatment and holdout arms, making the proportion of Quants among targeted users, $\Pi^{Targ}_{Q|A}$, the same for targeted exposed and unexposed users. Poets and Quants respond differently to the ad, so the incremental effect of exposure changes with the proportion of Quants who are targeted.

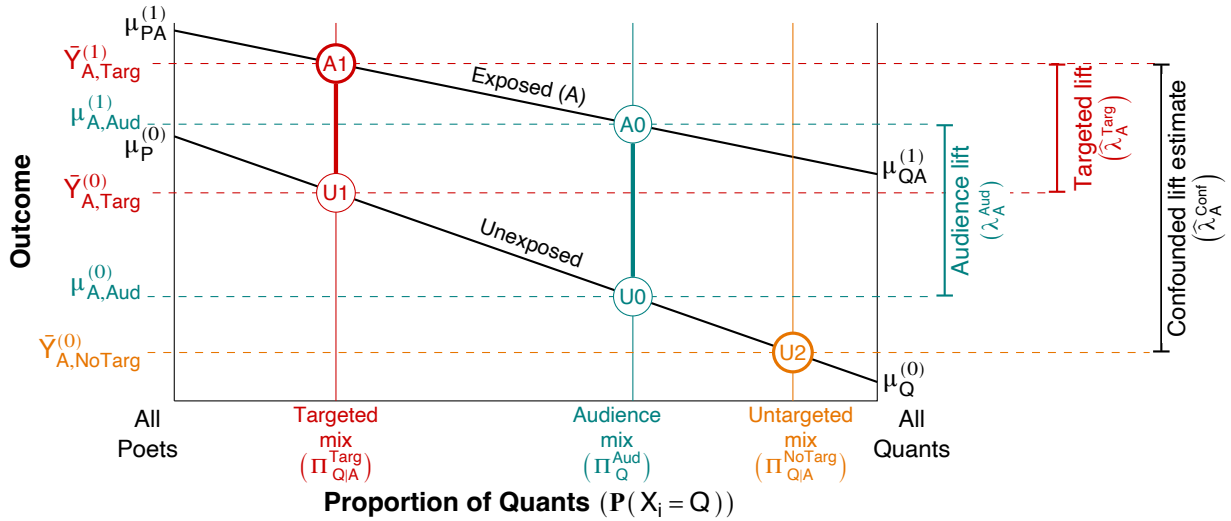
Figure 7 also shows the problems associated with comparing targeted exposed users to untargeted unexposed users, when those groups of users have different mixtures of user types. Let's consider what would happen if the platform did not offer the holdout tests that are necessary for $\widehat{\lambda}_{A}^{Targ}$ to be a truly causal estimate. That is, instead of reporting $\bar{Y}^{(0)}_{A,Targ}$, the platform instead reported $\bar{Y}^{(0)}_{A,NoTarg}$, computed from realized outcomes of untargeted unexposed users. In that case, the estimated lift

observed by the experimenter would be $\widehat{\lambda}_{A}^{Conf} = \bar{Y}_{A,Targ}^{(1)} - \bar{Y}_{A,NoTarg}^{(0)}$, which is a difference in average realized outcomes between targeted exposed users and untargeted unexposed users. In Figure 7, $\widehat{\lambda}_{A}^{Conf}$ is the vertical distance between (A1) and (U2).

This estimate *is not capturing the incremental effect of the ad creative*. Instead, $\widehat{\lambda}_{A}^{Conf}$ is a confounded estimate that captures two effects tangled together: the single-ad lift of the ad creative among targeted users (moving down in Figure 7 from (A1) to (U1)) and the comparison between the baseline estimated from the unexposed but targeted users and the baseline estimated from a completely different set of users who were not targeted with the ad (moving to the right from (U1) to (U2)). The confound arises because of the non-randomness of the exposure process. In essence, $\widehat{\lambda}_{A}^{Conf}$ is "contaminated" by combining two sets of users: the targeted exposed users who happen to have higher baseline outcome rates even if unexposed, and the untargeted users who have lower baseline outcome rates (and are always unexposed). As a result, this confounded estimate of the ad's lift overestimates the actual lift among the audience. To avoid this confound in targeted advertising environments, experimenters need to use properly randomized single-ad tests with holdout when computing lift, and such tools are already available.

For a single-ad lift using available ad A-B testing tools, experimenters can compute $\widehat{\lambda}_{A}^{Targ}$, which estimates the difference in outcomes between users targeted with and exposed to ad A to those who were also targeted with but randomly not exposed to ad A (vertical distance between (A1) and (U1) in Figure 7). However, the lift among the targeted users, $\widehat{\lambda}_{A}^{Targ}$, also does not properly estimate lift for the entire audience, λ_{A}^{Aud} (vertical distance between (A0) and (U0)). In this case, the computed lift among targeted users, $\widehat{\lambda}_{A}^{Targ}$, underestimates the true audience lift, λ_{A}^{Aud}, because: (1) Quants are more prevalent among untargeted than among targeted users ($\Pi_{Q|A}^{Targ} < \Pi_{Q}^{Aud}$); and (2) Poets are more likely to convert even when unexposed to the ad than Quants are when exposed to the ad ($\mu_{Q}^{(1)} < \mu_{P}^{(0)}$). The algorithm is targeting more users (Poets) who have a higher baseline conversion rate, rather than users with a higher potential *incremental* effect (Quants). Thus, the incremental aggregate lift is lower for the targeted users than for the audience overall.

Figure 7: An Ad Effect Depends on Baseline Outcomes and User Mixtures.



NOTE: Measuring lift of a single ad involves computing the vertical distance between intersection points. Intersections locate the expected or computed potential outcomes (y-axis) for a set of users with a given mixture (x-axis). Mixtures are shown by the proportions of Quants among subsets of users: the audience (green), targeted users (red), and untargeted users (orange). When the distance is *strictly* vertical between the Exposed and Unexposed lines, the effect is causal because exposed and unexposed users have the same type distribution. λ_A^{Aud} (the distance between (A0) and (U0)) is the effect of ad A in the audience, but this cannot be observed directly. In a holdout test, targeted exposed users and targeted unexposed users have the same type mix, so $\widehat{\lambda}_A^{Targ}$ (the distance between (A1) and (U1)) estimates the causal effect of ad A among targeted users. But if all exposed users are targeted and all unexposed users are untargeted (and outcomes from untargeted users are tracked), then $\widehat{\lambda}_A^{Conf}$ (the vertical distance between (A1) and (U2)) is confounded by the different mixtures of the two groups, so it is not a valid measure of lift.

A-B Differences, Divergent Delivery, and Sign Reversals

Returning our focus to comparisons between different ads, recall from Equation 2 that Δ_{AB}^{i} is the change in incremental outcomes caused by a single user's exposure to A, relative to exposure to B. It follows that $\Delta_{AB}^{X} = \lambda_{XA} - \lambda_{XB}$ is the expected value of Δ_{AB}^{i} among users with type X. And since the mix of user types in the audience is the same for users eligible for ads A and B (i.e., the partitioning of the audience is random), the A-B difference for the audience, Δ_{AB}^{Aud} is a mixture of Δ_{AB}^{X} over Π_{X}^{Aud}.

But under divergent delivery, the mix of users targeted with A differs from those targeted with B (Figures 2a and 3b). Unlike models of expected outcomes and lifts for *single ads*, there is no single type distribution over which Δ_{AB}^{X} can be mixed because targeted users who are eligible for each of the *two different ads* have different type distributions ($\Pi_{X|A}^{Targ} \neq \Pi_{X|B}^{Targ}$). Therefore, as experimenters estimate the A-B difference with $\widehat{\Delta}_{AB}^{Targ} = \widehat{\lambda}_A^{Targ} - \widehat{\lambda}_B^{Targ}$, they are comparing unbalanced mixtures of

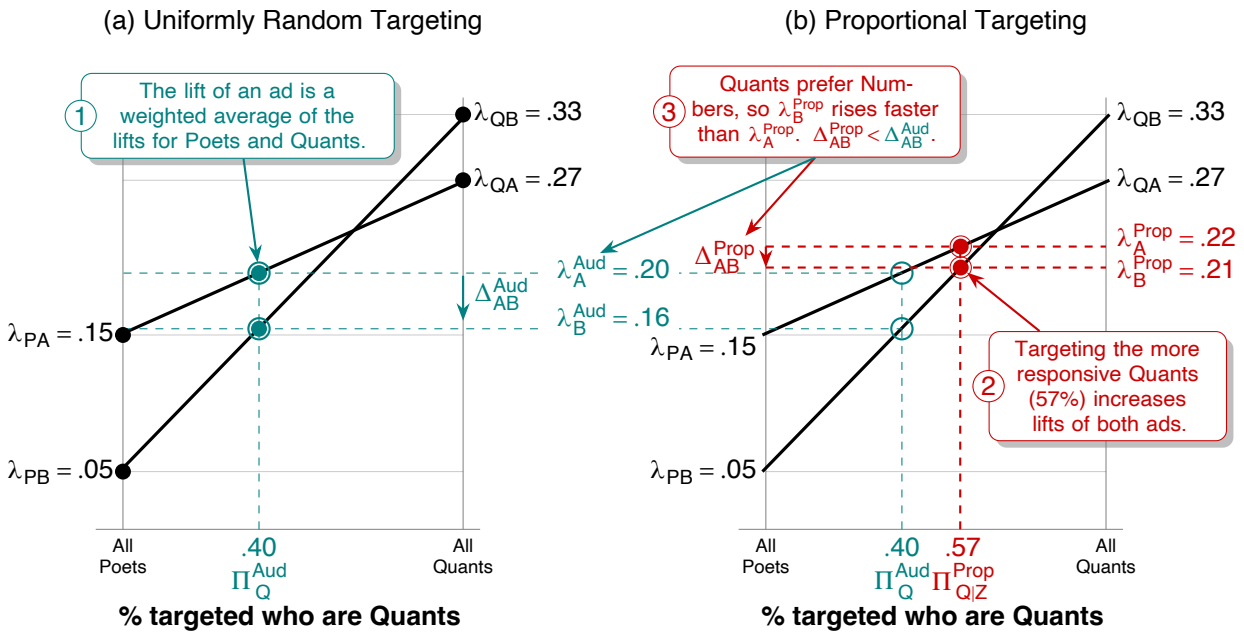
targeted users. That is, $\widehat{\Delta}_{AB}^{Targ}$ is confounded by how the algorithm targets different mixes of users eligible for A than users eligible for B. The experimenter cannot know how much of the $\widehat{\Delta}_{AB}^{Targ}$ is capturing the effect of ad content and how much is an artifact of targeting different types of users to be exposed to each ad. Web Appendix B provides a more formal treatment of these arguments.

Figures 8 and 9 help develop this intuition by illustrating how targeting policies and user heterogeneity conspire to confound A-B comparisons. To make the examples more concrete and easier to compare across panels, we describe A as the Words ad and B as the Numbers ad, and we apply numerical values to the lifts. Figure 8 describes two different targeting policies, neither of which deploys divergent delivery, but each yields different mix of heterogeneous targeted users. Figure 8a depicts uniformly random exposure to an audience, where exposed users are both representative of the audience and balanced between ads. Figure 8b shows a proportional targeting policy, which is non-representative (the proportion of Quants in the audience is 40%, while the proportion of Quants among all targeted users is 57%), but balanced (that targeted mix is the same for both ads). The estimates for lifts and A-B differences come from aggregating across each of the mixes, where each mix is determined by its targeting policies.

Figure 8a displays uniform random targeting that would generate data that reflect true effects of the ads in audience. But in Figure 8b, proportional targeting, the test targets more Quants (who are the better responders overall) than their incidence in the audience, so the experimenter will overestimate the effects of the ads relative to the true effect in the audience ($\lambda_A^{Prop} > \lambda_A^{Aud}$ and $\lambda_B^{Prop} > \lambda_B^{Aud}$). While the Quants are targeted equally for both ads, those Quants respond stronger to the Numbers ad than the Words ad, so the overestimation is not equal across ads; instead, it is greater for λ_B^{Prop} than for λ_A^{Prop}. Because users are heterogeneous in how they respond to the ads (Quants respond more than Poets to both ads), the estimated A-B difference among targeted users under proportional targeting is smaller than the true A-B difference in the audience ($\Delta_{AB}^{Prop} < \Delta_{AB}^{Aud}$). However, because the effects of the Words and Numbers ads are both computed from mixtures of users with the same proportion of Quants, the A-B comparison maintains its internal validity *within* the subset of targeted users.

Figure 9 illustrates how the estimated A-B difference changes under divergent delivery. The audience

Figure 8: Comparing Aggregated Lifts (No Divergent Delivery)

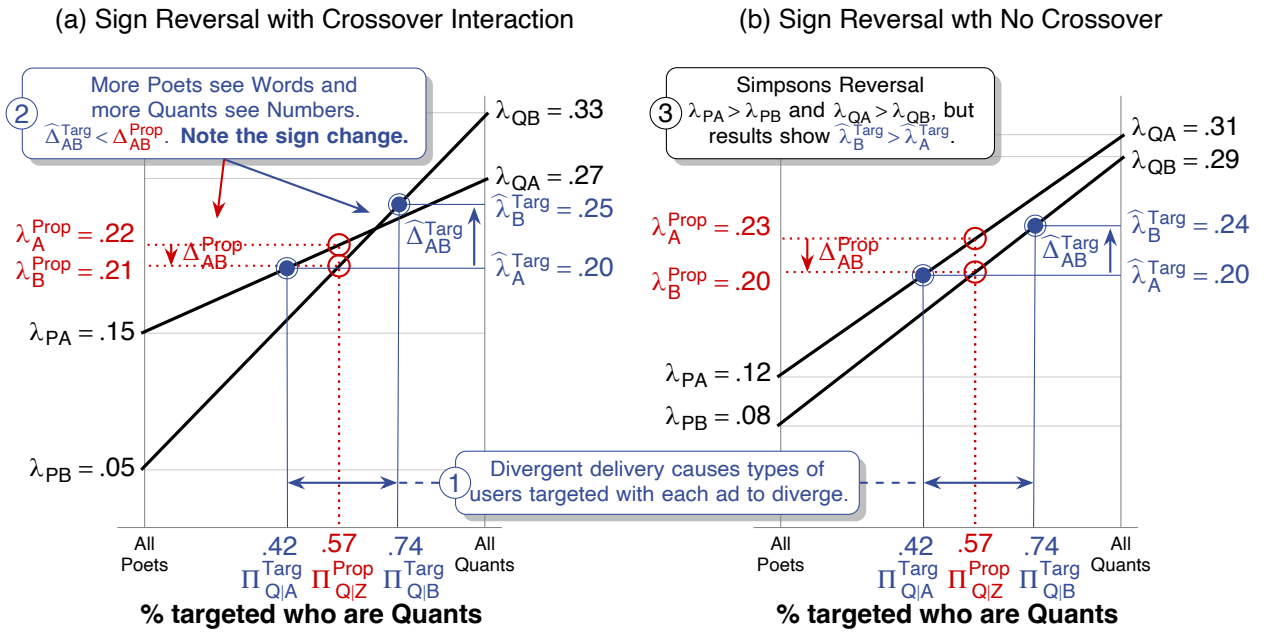


NOTE: Diagonal lines show lifts for Words (A) and Numbers (B) ads, aggregated over a mixture of Poets and Quants. In both panels, the mix of types in the audience is 60% Poets and 40% Quants, and users are equally likely, on average, to be targeted with either ad ($\Phi_{\bullet A} = \Phi_{\bullet B}$). The aggregate conversion rate is 15% higher for Words (A) than for Numbers (B) ($\mu_A^{(1)}/\mu_B^{(1)} = 1.15$) and three times higher for Quants than for Poets ($\mu_P^{(1)}/\mu_Q^{(1)} = .33$). The aggregate expected potential outcome is $\widetilde{\mu}^{(1)} = .2$, and we normalize the expected baseline conversion rate for unexposed users to $\mu_{XZ}^{(0)} = 0$. The quantities λ_A^{Prop}, λ_B^{Prop}, and Δ_{AB}^{Prop} are special cases of $\widehat{\lambda}_A^{Targ}$, $\widehat{\lambda}_B^{Targ}$, and $\widehat{\Delta}_{AB}^{Targ}$ that would have been observed if targeted were proportional (balanced). .

in Figure 9a has the same characteristics as in Figure 8b, but now there is divergent delivery, so the targeting probabilities for each type of user differ by ad (instead of the mix being the same across both ads). Consider the Numbers ad, which is targeted to a mix of users with 74% Quants (26% Poets). The targeting algorithm delivers the Numbers ad to more of the users (Quants) who respond better to that ad, making the estimated lift of Numbers, $\widehat{\lambda}_B^{Targ}$, higher under divergent delivery than it would have been under proportional targeting. The Words ad, on the other hand, is targeted to mix of users who are 42% Quants (58% Poets). Although Poets respond more to Words than Numbers, Poets are less responsive than Quants to both ads overall. The advertiser gets more of a bump in conversions from Quants seeing Numbers than from Poets seeing Words ad. As a result, divergent delivery pushes the estimated lift of Words below that of Numbers ($\widehat{\lambda}_A^{Targ} < \widehat{\lambda}_B^{Targ}$).

In this example, divergent delivery not only affects the magnitude of the estimated A-B difference, but it also causes the *sign* to flip. That estimate under divergent delivery is a reversal of what the

Figure 9: Sign Reversals in Aggregate Lifts under Divergent Delivery with and without Crossover Interactions in Response Rates.



NOTE: Quants are twice as likely to be targeted than Poets, but divergent delivery ($\rho_\tau = 4$) causes the mix of Quants to "diverge" by ad. The sign of the estimated A-B difference flips from the proportional targeting case (Figure 8b). Sign reversals can occur with (Figure 9a) and without (Figure 9b) crossover interactions in conversion rates. The right panel, without crossover interaction, shows a Simpson's Reversal: Words is more effective among both Poets and Quants, but the experimenter's estimate shows Numbers to be stronger.

experimenter would have observed under a balanced targeting policy. Based on Figure 9a, the experimenter will infer that Numbers-focused ad B is stronger than Words-focused ad A ($\widehat{\Delta}^{\text{Targ}}_{\text{AB}} = +.05$), even though identical mixes of users would have converted more when exposed to Words than Numbers ($\widehat{\Delta}^{\text{Prop}}_{\text{AB}} = -.01$). Importantly, *there is no way for the experimenter to know if the estimated sign from their* A-B *test results is different from what it would have been without divergent delivery.* This will be undetectable to the experimenter because they only observe the estimated *aggregate* effects, $\widehat{\lambda}^{\text{Targ}}_{\text{A}}$ and $\widehat{\lambda}^{\text{Targ}}_{\text{B}}$, but not the latent user-type-specific effects, λ_{XZ}, nor the mixture weights $\Pi^{\text{Targ}}_{\text{X|Z}}$. Experimenters should be concerned that different targeting policies lead to opposite inferences about how users respond to ad creatives.

Sign reversals are most likely to arise when the targeting algorithm is *overtargeting*: implementing a highly divergent targeting policy when the true differences in ad effects in subgroups are actually quite small. And this may be quite common in ad A-B testing because it is a key part of typical online

ad delivery. To effectively reach a heterogeneous mix of users, experimenters and platforms want to exploit even small differences in predicted responses to ads. If ads in an experiment share some common creative elements (e.g., positioning strategy, language, imagery, etc.), then it is plausible for the difference in the effects of those ads to be small.[17] In that case, experimenters should prefer the algorithm to be cautious about changing the mix of types targeted with each ad by too much during A-B tests. But if the algorithm uses creative elements as the basis for delivering Words almost solely to Poets and Numbers almost solely to Quants instead, then the divergent delivery policy will likely be more extreme than heterogeneity in effects should warrant. If that happens, then the mixes of users targeted to each ad could diverge enough to create a sign reversal that the experimenter cannot identify.

It is worrisome enough that divergent delivery can result in incorrectly estimating the sign of an A-B difference in a purportedly randomized experiment, compared to an A-B test without divergent delivery. But Figure 9b illustrates an even more concerning example of this sign reversal. Unlike Figure 9a, the users in Figure 9b do not exhibit a crossover interaction in how they respond to the two ads. Here, Words is truly the stronger ad among *both* Poets and Quants (and therefore, for any mixture of them). Yet to the experimenter, the reported A-B test results show Numbers as the stronger ad. Such an reversal effect stems from theoretical underpinnings discussed in the consumer psychology literature (Hutchinson et al. 2000). This particular reversal is an example of an undetectable *Simpson's reversal* (Blyth 1972; Pearl 2014). A Simpson's reversal occurs when the true effect of A is greater than the true effect of B for all user types separately, but the estimates incorrectly show that B is stronger than A when aggregated across unobserved user types. That is, Poets and Quants both respond better to A than to B, yet comparing the targeted mixes in aggregate, B performs better than A (i.e., if $\lambda_{PA} > \lambda_{PB}$ and $\lambda_{QA} > \lambda_{QB}$, but $\widehat{\lambda}_{A}^{Targ} < \widehat{\lambda}_{B}^{Targ}$). Such a Simpson's Reversal will occur when: (1) the amount by which the stronger ad's effect exceeds the weaker ad's lift among targeted users *within each user type* is sufficiently *small*; (2) the difference between user types for the *weaker* ad's effect is sufficiently large; and (3) the users responding better to the weaker ad are more prevalent among users targeted with that weaker ad than they are among users targeted with the stronger ad.

Sign reversals are more likely inside an A-B test than outside. In fact, the requirements of the

[17] Small effect sizes are consistent with large meta-analyses in online advertising (Johnson et al. 2017b).

experimental design in an A-B testing setting could force (or at least encourage) the algorithm toward targeting policies that make a sign reversal more likely than in a non-experimental setting. In a non-experimental campaign, a targeting algorithm that quickly suspects the Words ad is stronger than Numbers ad among both types of users might give up on the Numbers ad, targeting only users who were assigned to Words and not delivering the Numbers ad at all. But an experiment designed to compare Words and Numbers needs to expose at least some fraction users to Numbers, even though it may end up being weaker ad overall. The algorithm will then try to get as many conversions from Numbers (the weaker ad overall) as it can by targeting it to a mix with more Quants (the better responders to the weak Numbers ad). Since the Quants respond better than Poets to both ads, the Quants assigned to Numbers will still outperform the Poets assigned to Words, leading to the reversal.

A-B Differences Across Targeting Policies and Response Heterogeneity Patterns

Next, we use numerical simulation as a tool to delve deeper into how targeting policies lead to different inferences from the same audience, and how forms of heterogeneity in those audiences moderate those impacts on inferences. The objective of the simulation study is to untangle the factors that cause the estimated A-B differences under divergent delivery to deviate from the true A-B differences in the audience or among proportionally targeted users. Simulation lets us study the effects of three different forms of targeting policies (divergent delivery, proportional targeting, and uniform random audience exposure) under various patterns of heterogeneity in user responses. This allows us to exogenously manipulate whether targeting is representative (vs unrepresentative) and whether it is balanced (vs unbalanced). The simulation gives our study a level of control that exceeds what is possible given empirical data and current limitations of platforms' experimental tools.

Simulation Setup

We present the main aspects of the simulation here, with full details in Web Appendix C. Each user can have one of four type-ad combinations: $X \in \{P,Q\}$ (Poets and Quants) crossed with $Z \in \{A,B\}$ (Words and Numbers). All users with type X who are eligible to receive ad Z are targeted with probability Φ_{XZ}. If they are exposed to their assigned ad, they convert with probability $\mu_{XZ}^{(1)}$, and if unexposed, they convert with probability $\mu_{X}^{(0)}$. These probabilities constitute a "world" that specifies

the distributions of potential outcomes for users and targeting decisions, which in turn govern realized outcomes. For our setting, a world is determined by 4 targeting probabilities (Φ_{PA}, Φ_{PB}, Φ_{QA}, Φ_{QB}), 4 conversion probabilities when exposed to the ad ($\mu^{(1)}_{PA}$, $\mu^{(1)}_{PB}$, $\mu^{(1)}_{QA}$, $\mu^{(1)}_{QB}$), and 2 baseline conversion probabilities when unexposed ($\mu^{(0)}_{P}$, $\mu^{(0)}_{Q}$). For each world, we generated potential outcomes for 15 simulated audiences of 200,000 users each. The users in these simulated audiences are 60% Poets and 40% Quants, and are assigned to be eligible to receive either Words or Numbers with equal probability.

To simulate an A-B test with holdout, we generated the full set of potential outcomes for each user, and targeted users under various policies to determine which of those potential outcomes for each user is actually realized. From those realized outcomes, we computed three sets of lifts and A-B differences: (1) estimated $\widehat{\lambda}^{\text{Targ}}_{A}$, $\widehat{\lambda}^{\text{Targ}}_{B}$, and $\widehat{\Delta}^{\text{Targ}}_{AB}$ that mimic what an experimenter would get from a "real world" A-B test conducted with divergent delivery (unrepresentative and unbalanced); (2) counterfactual $\lambda^{\text{Prop}}_{A}$, $\lambda^{\text{Prop}}_{B}$, and $\Delta^{\text{Prop}}_{AB}$ that an experimenter *might* have estimated had targeting during the experiment been proportional (unrepresentative, but balanced); and (3) counterfactual λ^{Aud}_{A}, λ^{Aud}_{B}, and Δ^{Aud}_{AB} from the simulated potential outcomes of all users in the audience (representative and balanced).

The simulation study manipulates the relationships among targeting probabilities and response rates. The algorithm targets Poets and Quants with a ratio ($\Phi_{P\bullet}/\Phi_{Q\bullet}$) that varies continuously from 1/5 to 5, and with one of five levels of divergent delivery, $\rho_{\tau} \in \{1/8, 1/4, 1, 4, 8\}$. In all conditions, users eligible to receive Words or Numbers are equally likely to be targeted. And users exposed to Words are *equally* likely, *15% more* likely, or *twice* as likely to convert than those who respond to Numbers.

But the degree of *response heterogeneity* across users is a moderating factor that can strengthen or attenuate the impact of targeting policies on A-B test results. We allow aggregate response probabilities to be either *equal* for Poets and Quants, or Quants are *3x* as likely to convert than Poets ($\mu^{(1)}_{P}/\mu^{(1)}_{Q} \in \{1,1/3\}$). We also allow for a *user-ad interacton* in response heterogeneity. With *proportional response*, the Poet-to-Quant ratio of expected conversions is the same for users exposed to Words or Numbers. But with *divergent response*, Poets convert more after exposure to one ad (say, the Words ad), and Quants convert more after exposure to the other ad (say, Numbers), relative to whatever the marginal conversion rates across ads and types alone would have dictated. Divergent

response yields a crossover interaction (user types and ads), as seen in Figures 8a, 8b, and 9a. By contrast, proportional response yields the pattern of parallel lines, as seen in Figure 9b, making it susceptible to a Simpson's Reversal. We define proportional and divergent response to be analogous to proportional and divergent delivery, but these characterize patterns of user responsiveness to ads.

Simulation Results

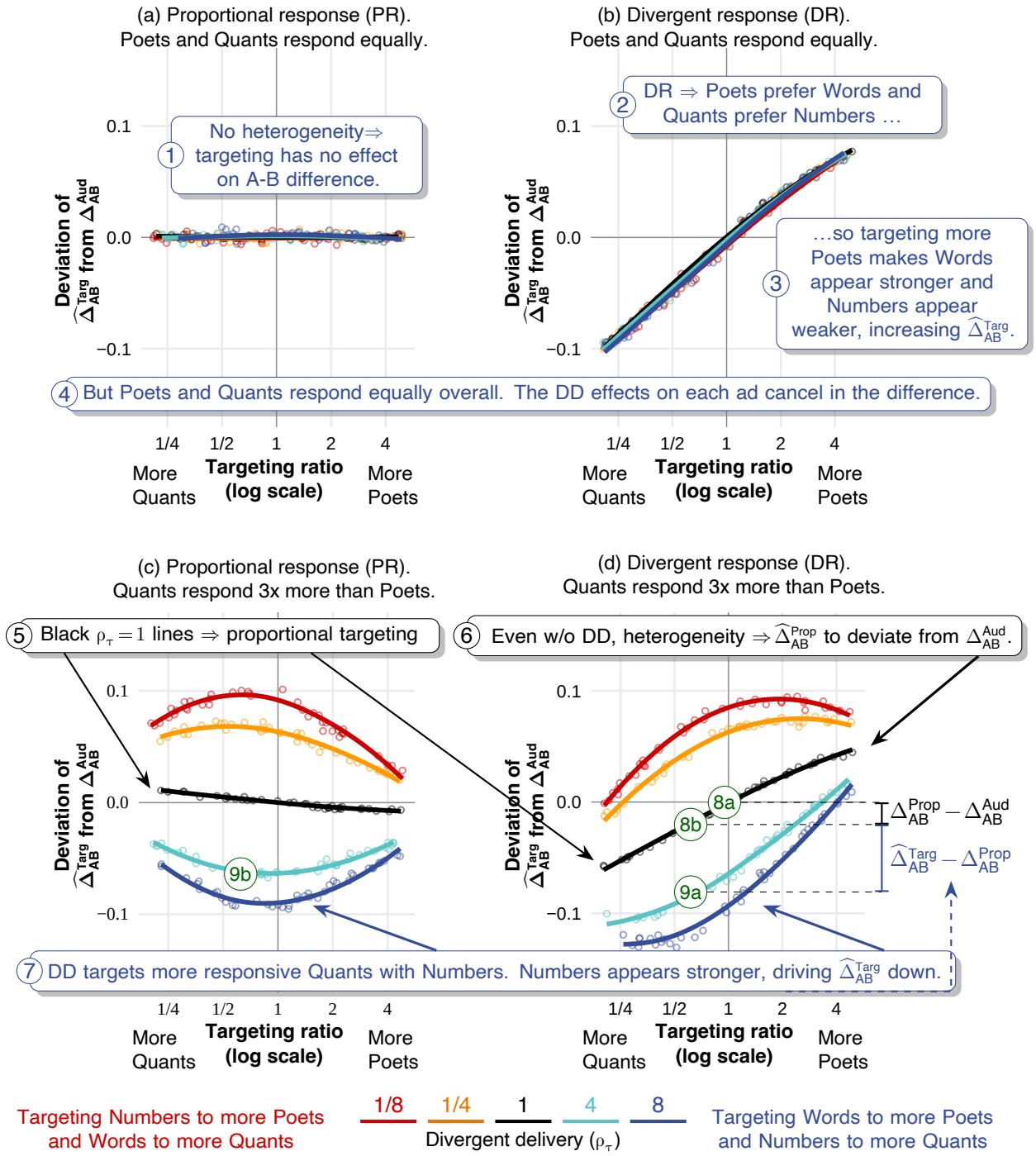
The first part of our simulation results compares the estimated $\widehat{\Delta}_{AB}^{Targ}$ (with a particular targeting policy) to two different baselines: Δ_{AB}^{Aud} (which would be the A-B difference under uniformly random targeting), and $\widehat{\Delta}_{AB}^{Prop}$ (under proportional targeting). In Figure 10, each point represents an average across the 15 audiences that were simulated under the same "world" of parameters.

Variation across panels correspond to differents pattern of audience response heterogeneity: rows differ by the comparative aggregate response heterogeneity between Poets and Quants, and columns differ by whether users exhibit proportional response or divergent response. In all panels, the aggregate response rate for Words is 15% greater than for Numbers.

Variation within each panel in Figure 10 describes targeting policies in terms of two dimensions: marginal targeting ($\Phi_{P\bullet}/\Phi_{Q\bullet}$) along the x-axis and divergent delivery (ρ_τ) with color. When $\rho_\tau > 1$, the targeting policy includes divergent delivery favoring Poets to see Words and Quants to see Numbers. Similarly, $\rho_\tau < 1$ refers to targeting with divergent delivery favoring Quants exposed to Words and Poets exposed to Numbers. Considering targeting policies and response heterogeneity *jointly* is important. For instance, when the audience has divergent response favoring Poets exposed to Words and Quants exposed to Numbers, a divergent delivery targeting policy with $\rho_\tau < 1$ is "mistargeting" ads to users, since users would be less likely to be exposed to their preferred ads.

The y-axes in Figure 10 shows how much $\widehat{\Delta}_{AB}^{Targ}$ with divergent delivery deviates from the theoretical Δ_{AB}^{Aud} for the audience. But we can split this deviation into two components. First, the black line is the benchmark case of proportional targeting ($\rho_\tau = 1$, no divergent delivery), where the mixes of targeted Poets to Quants are balanced across ads. This represents a difference equal to $\Delta_{AB}^{Prop} - \Delta_{AB}^{Aud}$, which is the impact of losing representativeness under proportional targeting relative to uniform random exposure. Second, the amount of the deviation attributable to divergent delivery, $\widehat{\Delta}_{AB}^{Targ} - \Delta_{AB}^{Prop}$, is the

Figure 10: A-B differences by Heterogeneity Pattern and Targeting Policy



NOTE: Follow the numbered signposts in order. Each panel represents a pattern of response heterogeneity. Targeting policies are on the x-axis (proportion of targeted Poets to Quants, log scale) and color (degree of divergent delivery, "DD"). The y-axis shows the difference between $\widehat{\Delta}^{Targ}_{AB}$ and Δ^{Aud}_{AB}. The isolated effect of divergent delivery is the vertical distance between any colored line (ρ_τ) and the black line ($\rho_\tau = 1$). For targeting policies using divergent delivery with $\rho_\tau > 1$ (light and dark blue curves), Poets are more likely to be targeted with Words, and Quants, more with Numbers. When divergent delivery is in the opposite direction with $\rho_\tau < 1$ (orange and red curves), then Poets are more likely to be targeted when eligible to receive Numbers, and Quants, with Words. The combinations of heterogeneity and targeting policies from Figures 8 and 9 are at the circled points (8a), (8b), (9a), and (9b).

vertical distance between any colored line with divergent delivery ($\rho_\tau \neq 1$) and the black line with proportional targeting ($\rho_\tau = 1$).

We discuss the patterns of these deviations for each panel in turn. In Figure 10a, the worlds exhibit no response heterogeneity. With no variation in how different users respond when exposed to each ad, targeting policies do not affect aggregate results of the A-B test, so $\widehat{\Delta}^{\text{Targ}}_{\text{AB}} = \Delta^{\text{Prop}}_{\text{AB}} = \Delta^{\text{Aud}}_{\text{AB}}$.

In Figure 10b, Poets and Quants have the same marginal conversion rates across ads. Words is more effective than Numbers, in aggregate across users (for all panels in Figure 10). And divergent response means that the conversions from Poets are even more likely to result from exposure to Words, and the Quants' conversions are more likely to come from their exposure to Numbers. Also, targeting more Poets overall increases the gap in estimated A-B differences. Moving left-to-right along the x-axis means targeting more Poets, which skews the mix of users exposed to Words in favor of that ad's best responders. It also skews the mix of users exposed to Numbers away from Quants, its best responders. Therefore, $\widehat{\lambda}^{\text{Targ}}_{\text{A}} > \lambda^{\text{Aud}}_{\text{A}}$ is overestimating the effect of Words and $\widehat{\lambda}^{\text{Targ}}_{\text{B}} < \lambda^{\text{Aud}}_{\text{B}}$ is underestimating the effect of Numbers. The deviation of $\widehat{\Delta}^{\text{Targ}}_{\text{AB}}$ from $\Delta^{\text{Aud}}_{\text{AB}}$ goes up.

But despite the effect of marginal targeting of users, divergent delivery has no effect on the A-B difference beyond effect of proportional targeting here (the ρ_τ lines are colinear but not flat in Figure 10b). Because the marginal conversion rates of Poets and Quants are equal, divergent delivery has an equal and opposite effect on each ad separately. For instance, a divergent delivery policy with $\rho_\tau > 1$ targets more Poets assigned to Words, further increasing the estimated effect of Words. It also targets more Quants assigned to Numbers, increasing the estimated effect of Numbers. Therefore, in this case, the effects on the separate ads cancel out when taking the difference between A and B.

In Figure 10c, the marginal conversion probability for Quants is 3 times that for Poets, and that ratio is the same for both ads (proportional response). Without divergent delivery ($\rho_\tau = 1$, black line) targeting more Poets still only slightly decreases the estimated effect of both ads. But those effects cancel in the A-B difference (black line is nearly flat). The slight downward slope arises because Words is stronger than Numbers, and there are more Poets than Quants in the audience to begin with.[18]

[18] In the simulation, $\Pi^{\text{Aud}}_{\text{Q}} = .4$. This line would be flat for $\Pi^{\text{Aud}}_{\text{Q}} = .5$.

Therefore, $\widehat{\lambda}_{A}^{Targ}$ is more sensitive than $\widehat{\lambda}_{B}^{Targ}$ to targeting more Poets. But targeting with divergent delivery ($\rho_\tau > 1$, blue lines) targets even more of the less responsive Poets with Words and more of the more responsive Quants with Numbers, creating a negative shift in $\widehat{\Delta}_{AB}^{Targ}$away from Δ_{AB}^{Aud} and Δ_{AB}^{Prop}.[19]

In Figure 10d, while Quants are also 3 times as likely to convert than Poets, that ratio is smaller for Words and higher for Numbers (divergent response). Even without divergent delivery, the decline in estimated lift of Numbers, $\widehat{\lambda}_{B}^{Targ}$, caused by targeting fewer of its best responders (Quants), is more than offset by the increase in estimated lift of Words, $\widehat{\lambda}_{A}^{Targ}$, by targeting more of its best responders (Poets) with Words. Thus, $\widehat{\Delta}_{AB}^{Targ}$ increases as more Poets are targeted, even without divergent delivery. Divergent delivery with $\rho_\tau > 1$ still targets fewer of the less responsive users (Poets) and more of the more responsive ones (Quants). But the magnitude of the deviation is influenced by the relative strengths of the ads for each type of user.

In Figs. 10c and 10d, the numbers in the circled points (8a), (8b), (9a), and (9b) refer to the particular combination of audiences and targeting policies that are reflected in the corresponding numbered subfigures in Figures 8 and 9. Taken together, the progression from (8a) to (8b) to (9a) illustrates how $\widehat{\Delta}_{AB}^{Targ} - \Delta_{AB}^{Aud}$ can be decomposed into two components. The first component captures proportional vs. uniformly random targeting, $\Delta_{AB}^{Prop} - \Delta_{AB}^{Aud}$ (the vertical distance from (8a) to (8b) in Figure 10d), is the effect of targeting more Quants than Poets, but in equal proportions, relative to uniformly random exposure. Considering Figure 8a and Figure 8b, we hold the audience constant but observe two different targeting policies. The second component, $\widehat{\Delta}_{AB}^{Targ} - \Delta_{AB}^{Prop}$ (the vertical distance from (8b) to (9a) in Figure 10d), is the incremental effect of divergent delivery above and beyond the effect of proportional – not uniformly random – targeting. Those two circled points, also reflect two examples of reversals, shown in Figs. 9a and 9b, which have the same level of divergent delivery but differ by response heterogeneity.

In summary, the simulation results show divergent delivery has larger effects on the estimate of A-B differences when some users respond more than others and when that heterogeneity also exhibits

[19]The effect of divergent delivery diminishes at the extremes of the x-axis, where the marginal targeting policy is already targeting nearly all Poets or Quants. If nearly all targeted users are already of one type, then there is little that divergent delivery alone could do to change that mixture.

divergent response. But when different users respond similarly to any given ad, the impact of divergent delivery is reduced.

Ecological Consequences of Divergent Delivery

When and for Whom Does Divergent Delivery Matter Most?

Not all experimenters need to learn the same thing from A-B tests. How much value one extracts from a test conducted on targeted ads comes down to whether the experimenter needs (or even wants) to isolate the causal effect of the content of the ads from the effect of the targeting algorithm (balance), and whether comparisons need to be made within the scope of the defined audience (representativeness).

Consider the marketer who conducts A-B tests solely to predict which creative elements are likely to "perform better" during the rollout of a ad campaign (say, an ad buyer for a digital marketing agency). The success of the campaign depends on the bundled value proposition of the platform: the advertiser is buying not only "eyeballs" for their ads, but also the use of the algorithm to select the "right eyeballs" for each ad separately. This experimenter needs the test to replicate the conditions of the rollout, which will occur in the presence of divergent delivery. For this use case, the unbalanced exposure to ad treatments that is induced by divergent delivery could actually be desirable, but only if the algorithm itself does not change between the test and rollout phases. More generally, even assuming the platform's algorithm is stable may be problematic given that it learns about which users respond to different creative content over time.

Divergent delivery *is* a problem for experimenters who want to make causal inferences about any "active ingredient" in their ads. This is the goal of the landscaper we introduced earlier, who is comparing user responses to different creative elements to develop marketing strategy and inform positioning of the brand. Meeting these objectives involves drawing insights that are ecologically valid beyond the exact conditions of the test (e.g., a particular platform or online format). Thus, estimates of the effect of ad content need to be untangled from how the algorithm determines which users are targeted with each ad. This causal inference can only happen with balanced experimental designs,

which do not happen under divergent delivery. Due to that divergent delivery, even researchers internal to platforms are unable to make causal inferences from their own A-B tests (Gordon et al. 2023).

Extrapolating inferences from online A-B tests to support offline and strategic marketing decisions is something marketers are doing in practice, which exposes them to the pitfalls of experimentation under divergent delivery. For instance, the analytics team of one Fortune 500 company explained the extreme challenges of running randomized controlled field experiments that compare performance of different ad creatives on television. So instead, they test those elements of ad creatives in online A-B tests, and apply those results to a broader set of advertising decisions, including traditional channels. As another example, a marketing manager for a major US transportation company explained their decision-making process as "the quantitative informs the qualitative." For these marketers, A-B testing ads to a stock of knowledge that can be applied later. Even if the immediate objective of an A-B test is not directly connected to a strategic decision, the results of the test contributes to the company's understanding of its customers' preferences and behaviors. Additionally, a researcher who previously worked for a major online ad platform told us that members of its sales organization often interact with advertisers who are applying A-B test results beyond the scope of the ad platform itself, and who are not aware that the results cannot take a causal interpretation of the comparative effect of ad creatives.

Put simply, divergent delivery creates risk that strategic marketing and tactical decisions across the marketing mix (e.g., online ads elsewhere, offline promotion, brand positioning) would be based on confounded test results. Acting on these test results may be quite costly, especially if the direction of the estimated effect of the active ingredient of an ad is different than what would have been observed under proportional targeting where the different ads were delivered to a comparable mix of users.

Whether the subject pool of the test needs to be representative of the audience is a separate question from the need for balance. Because the targeting process relies in part on the relevance of ad content to users, A-B test subjects form a non-random subset of the audience. This non-representativeness during a test should be desirable for experimenters who are predicting ad *performance* during a future rollout phase.

But tests of causal effects of ad content may also benefit from a targeted subject pool that is pur-

posefully not representative of the advertiser-defined audience. The audience is defined along rather coarse criteria that are limited to observable characteristics. Tt may often be too broad to reflect an "optimal" market segment for which the product and brand should be positioned. As a result, experimenter might prefer an subject pool consisting of "whichever users the algorithm decides to target," even if it is unrepresentative of the audience. This is helpful when the experimenter's objective is to inform strategic marketing decisions. The question of representativeness is independent of the balance of the mix of users exposed to each ads. Even if targeting were proportional (with balanced exposure of users to each ad), and the goal of the experiment is to test the effects of creative elements in isolation from the targeting algorithm, marketers may want to trust the algorithm to run the balanced test on users for whom the brand is most relevant.

Using that non-representative target for the A-B test, however, involves a tradeoff. The downside of letting the algorithm choose the subject pool is that the experimenter would not be able to describe the distinction between the population of the focal market segment and the population on which the A-B test was conducted. But this lack of representativeness does not threaten the internal validity of the A-B comparison. Experimenters may be willing to live with not knowing the precise description of the A-B test subject pool because they trust the targeting algorithm to do a good job finding users relevant to the tested ads. That may be well worthwhile to the experimenter, as long as, among those targeted relevant users, there is a balanced randomized test that will be run.

Academics Using A-B Testing Tools Face An Additional Pitfall

An example where the experimenter likely does prefer both balance and representativeness is the case of the "academic researcher," whose A-B tests on targeted ad platforms are field experiments that test hypotheses about human behavior in settings that are more realistic than a lab. Their goal is to learn insights from the A-B tests with the same strength of evidence that a rigorous randomized controlled trial carries. Unlike experiments in a typical randomized control trial, the experimenter in an online ad A-B test does not have direct control over randomizing or observing participants. Therefore, the academic researcher needs more control over which users are included in the test than platforms' A-B testing tools currently provide. Because targeted users are not representative of the

predefined audience, the academic researcher cannot explain precisely why some users were included and others were not. Even more pernicious, the lack of balance across treatment conditions caused by divergent delivery means the researcher cannot make causal comparisons of ad content, even among the users targeted for that test. These A-B tests results would not carry the same evidentiary weight as a randomized controlled trial.

Still, the use of online A-B testing tools has become standard practice in consumer behavior research. For example, published studies in Orazi and Johnston (2020), Kupor and Laurin (2020), and Banker and Park (2020) manipulate creative elements of ads on Meta (Facebook/Instagram), and Cecere et al. (2018) randomize users to see different ads using Snapchat. In fact, the *Journal of Marketing* has published several papers that describing at least one experiment using a targeted advertising platform, and that draw inferences as if those experiments were properly randomized (Paharia 2020; Paharia and Swaminathan 2019; Winterich et al. 2019; Mookerjee et al. 2021; Zhou et al. 2022; Atalay et al. 2023) We understand that the authors who used platforms' A-B testing tools were following the best advice at the time.[20] However, if the A-B tests reported in these papers were meant to represent evidence of causal A-B comparisons across ad creatives, then divergent delivery rendered the A-B testing tools used for those tests inappropriate for those studies (Braun et al. 2024).

This is not to say that all A-B tests for academic research must be representative and balanced. Braun et al. (2024) describe situations where academics studying consumer behavior may want to sacrifice representativeness or balance to focus on effects in the context of a modern advertising environment where ad content and targeting policies are confounded "in the field." But balance is required as long as the goal is to test psychological constructs that are operationalized by *ad content*, isolated from the effect of that content from the confounding influence of targeting. If such psychological constructs are the focus, then the academic researcher should avoid causal claims from A-B tests that compare ad creatives. Given the range of the goals of academics who may use A-B testing tools, understanding when these methods are appropriate is crucial not only to other academics reviewing or building on consumer research findings, but *also to practitioners who rely on academic research to inform their own decision-making*.

[20] We did the same thing with our Detroit A-B test. The lack of balance across treatments motivated us to write this paper.

Dispelling Myths About Divergent Delivery

The pervasiveness of A-B tests among practitioners and academics has caused some confusion regarding the issues we raise in this paper, which we now take an opportunity to resolve.

Holdout Tests Do Not Resolve the Divergent Delivery Problem in A-B Comparisons

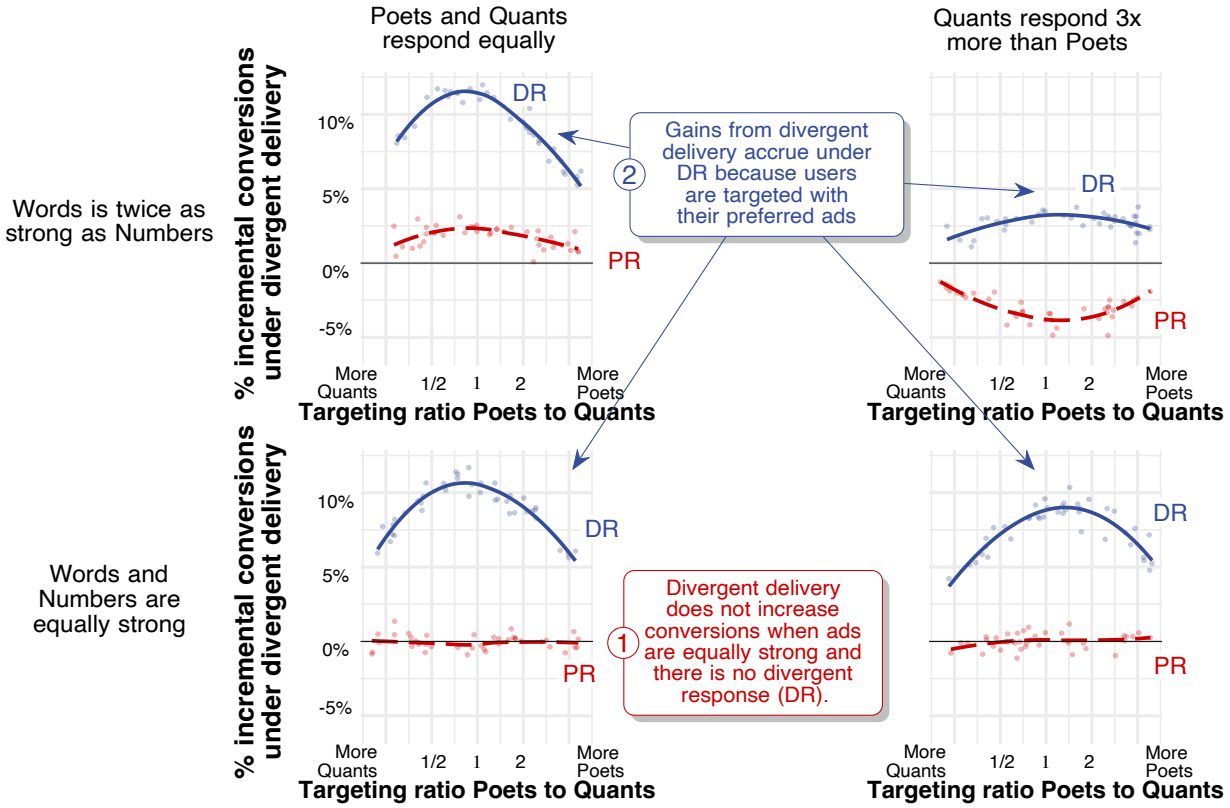
The term A-B test has sometimes referred to any online experiment where users are assigned to different experimental conditions, such as the treatment and holdout arms in a single ad holdout test (Figure 3a). Because assignment to these arms is random, some may think that comparing ads using A-B tests with holdouts solve the internal validity problems caused by of divergent delivery. Figure 3b illustrates why this is simply not the case. When targeted users are unbalanced across ads, splitting those users into treatment and holdout arms is no help.

Platforms Cannot and Will Not Disable Divergent Delivery

Given the challenges posed by divergent delivery for A-B tests, one might be curious about the implications of a counterfactual situation where balanced A-B tests could be conducted with divergent delivery "disabled." But experimenters should not expect platforms to make it possible to run A-B tests without divergent delivery anytime soon.

Divergent delivery is a profitable feature of the targeting algorithm, and disabling it incurs an opportunity cost. Figure 11 extends our simulation study to show the expected percent of incremental conversions resulting from a divergent delivery targeting policy compared to proportional targeting. Divergent delivery is more effective in the presense of divergent response (DR; blue curves) than proportional response (PR; red curves). So when Poets who are exposed to Words and Quants who are exposed to Numbers are most responsive, targeting with that same pattern generates more incremental conversions. Only if ads were equally effective (bottom row) and there were no divergent response across user types (red curves) would divergent delivery would have no incremental benefit. In essense, *disabling divergent delivery raises the short-term cost of running the experiment*. The economic value from divergent delivery can explain the market equilibrium where platforms do not offer an option to disable divergent delivery during experiments, and where experimenters accept this.

Figure 11: Percent Incremental Conversions from a Divergent Delivery Targeting Policy



NOTE: The y-axis shows percent incremental conversions under a divergent delivery targeting policy with $\rho_\tau = 4$, relative to proportional targeting with $\rho_\tau = 1$. Panels vary by heterogeneity in marginal responses across ads (rows) and user types (columns). Line colors indicate either proportional response (PR, red) or divergent response (DR, blue). The x-axes are the same as in Figure 10: how many more Poets or Quants are targeted across both ads.

One might then consider whether the platform might give experimenters the option to disable divergent delivery for an additional fee. But the business model of the platform is fundamentally misaligned with those of experimenters objectives (de Langhe and Puntoni 2021). Helping marketers make general strategic marketing decisions is not part of the platform's business. The platform does not want to make it easy to extrapolate information gleaned from A-B tests on their platform to be used to develop creative material for competitors' platforms or offline advertising channels. Thus, there is no incentive to allow experimenters to constrain A-B tests to be balanced.

Also, the details of how the algorithm targets users is the platform's proprietary "secret sauce." Letting experimenters compare A-B test results with and without divergent delivery enabled would allow experimenters to separately parse out the effect of their ads and the effect of the targeting algorithm's

user selection, effectively disassembling the platform's bundled value proposition. The operation of an online advertising platform depends on the integrated performance of many different components: the bidding algorithm, user classification, determining ad relevance, etc. The incremental value of divergent delivery in a targeted ad platform is not something the platform wants to make easy to "reverse engineer" and compare against competitive offerings.

Even if the platform did have a business incentive to allow for disabling divergent delivery, it may not even be possible. Targeting ads to users is so central to the value a platform offers advertisers that consideration of ad relevance is utterly intertwined in how the platform operates (Gordon et al. 2023). It should not be surprising that disabling even one component, buried deep in the inner workings underlying the largest advertising marketplaces in the world, is a difficult task.[21]

Researchers Unsuccessfully Attempt To Eliminate the Divergent Delivery Confound

The algorithm cannot be "tricked" by experimenters seeking to avoid divergent delivery during their A-B tests. However, there remains some confusion about this in the literature. In particular, some have suggested an experimenter might be able to attain balance in an A-B test by selecting certain options when configuring the experiment. For instance, Orazi and Johnston (2020) write that an experimenter using the Meta platform's testing tools can prevent divergent delivery by setting the experimental objective to "Reach," instead of "Conversions." But that appears to be contradicted by both the Meta documentation about relevance in ad delivery (see Footnote 2) and Orazi and Johnston's own results, which show a lack of balance across observed demographic groups.[22] Because targeting algorithms take user-ad relevance into account when exposing users to ads during an experiment, claims that certain configurations of A-B tests can ensure balance across treatments are, at best, conjecture.

Similarly, others may wonder if defining separate audiences for separate campaigns and separate A-B tests along different observable moderating variables might be an effective workaround. This was the motivation for strategy attempted by Matz et al. (2017) and critiqued by Eckles et al. (2018). But the

[21] In private correspondence, an insider of a major platform recalled being told that disabling divergent delivery for A-B tests would require "two to three years of engineering work." Removing targeting policies is not as easy as flipping a switch.

[22] Eckles: https://bit.ly/DEonOJ

moment an experimenter runs ads outside of one A-B test (e.g., running separate A-B tests), there is no guarantee that users will see at most one treatment ad. Indeed, if two campaigns with identical ads are run, even if the audience definitions seemingly do not overlap, it is possible that the targeting algorithm will deliver them to overlapping groups of users. This feature is by design; one application of Meta's testing tools is to compare performance of ads across different audience definitions.[23]

General Discussion

An A-B test may appear to many marketers to be an easy way to run field experiments to learn about the effects of ad copy, imagery, and messaging. But experimenters who run A-B tests in targeted online advertising environments ought to know what they are really getting. The concern about what experimenters can learn from A-B comparisons stems from how online ad experiments are not like typical randomized controlled experiments. By using platforms' A-B testing tools, experimenters lose control over subject selection and treatment assignment when using platforms' A-B testing tools has attracted interest across disciplines (Johnson 2023). While some readers may already know about some of these issues, others may be surprised. The evidence of that surprise is the rapid increase in academic publications presenting these A-B tests as ideal randomized field experiments. But there are persistent misunderstandings among practitioners about what the test results actually measure. Our concern is not the mere usage of non-random unbalanced exposure A-B tests, but rather the *presentation of results as if they came from balanced experiments* and subsequent conclusions and managerial decisions based on those results. Depending on the experimenter's objectives, confounding of these effects by the targeting algorithm can lead marketers to make suboptimal strategic decisions.

Contributions To the Literature

Our warnings about divergent delivery during A-B tests are distinct from concerns previously expressed in the literature. Eckles et al. (2018) uncover divergent delivery patterns in the data used by Matz et al. (2017), who ran different ad treatments on Facebook as separate, simultaneous ad

[23] https://bit.ly/MetaVars

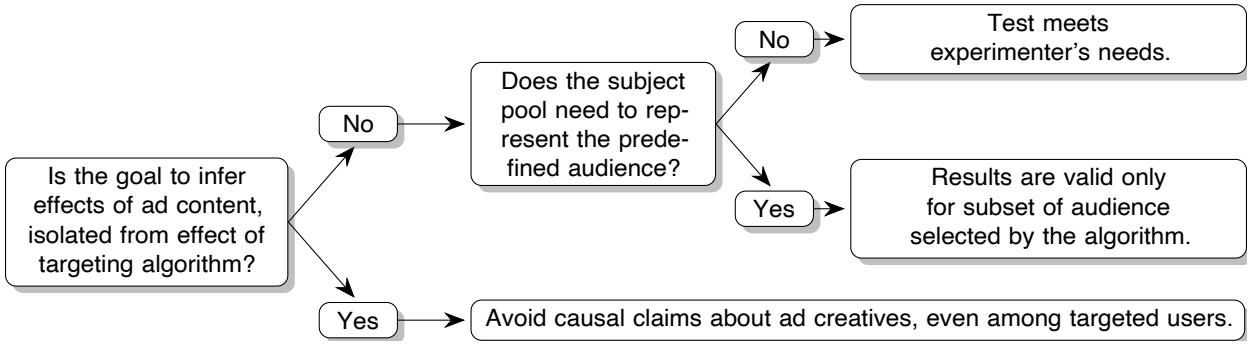
campaigns. That is, Matz et al. did not run their studies through a formal randomized experimentation tool. Eckles et al. argue that when an researcher *does not* run an A-B test through the platform, and instead runs multiple unlinked ad campaigns, then the internal validity of an A-B comparison is threatened. Therefore, they argue that without an A-B test, the researcher *should not* assume balanced randomization of users across ad treatments. We agree.

But this paper makes an additional, even stronger claim: even when an experimenter *does* run an A-B test through the platform, they *still should not expect users to be randomly assigned to ad treatments*. The platform is simply not randomly assigning even the targeted users to different ads. When relevance affects which users see which ads, and when results are aggregated across unobservable factors that the algorithm uses to assess that user-ad relevance, there is no way for the experimenter to separate how much of the A-B difference is caused by the creative elements of ads from how much of the effect is caused by the targeting algorithm. This paper is the first to provide evidence of and analysis of how divergent delivery occurs during an A-B test.

Further still, the problem caused by *divergent delivery remains even when comparing lifts from holdout tests of different ads*. While divergent delivery has been described in non-experimental settings of online advertising (Ali et al. 2019; Eckles et al. 2018; Lambrecht and Tucker 2019), and in experiments with placebo control ads (Barajas et al. 2016; Johnson et al. 2017a), it had not been shown to occur even when using the popular ad A-B testing tools comparing two or more ads, as recommended by Eckles et al. (2018) and Gordon et al. (2019). Our empirical evidence of divergent delivery, even across ads in A-B tests with holdouts, is the first to fill this gap in the literature.

This paper also adds to the literature by formally defining divergent delivery and its consequences. To investigate the effects of targeting and divergent delivery on aggregated results, we build on the potential outcomes model to distinguish *assignment* to multiple treatments from *exposure* to the assigned treatment. Our formal definitions of divergent delivery and response heterogeneity as conditional probabilities, odds ratios, and aggregated expected outcomes serves to structure discussion of the pitfalls surrounding experiments in targeted online advertising environments and the moderators of their impact on A-B test results.

Figure 12: What Should Experimenters Do About A-B Tests That Use Targeting with Divergent Delivery?



NOTE: Under divergent delivery, results must be interpreted as the combined effect of ad content and targeting. Whether this matters to the experimenter depends on the goal of the experiment.

Another contribution that distinguishes our paper from earlier work is that its takes the point of view of the experimenter, not the website or platform. Considering the experimenter's perspective lets us focus on the issue of *data aggregation*. The aggregation of outcomes across unobserved user types creates this key deviation that the experimenter cannot detect: the gap between the A-B comparisons estimated under divergent delivery and A-B comparisons that would have occurred under a balanced test. That balanced test could be either proportional targeting (balanced but unrepresentative) or uniformly random (balanced and representative) targeting. Since experimenters who are trying to make causal inferences about ad content may still prefer the subject pool to be selected by the targeting algorithm (balanced but unrepresentative), we think the deviation from the proportional targeting baseline ($\widehat{\Delta}^{\text{Targ}}_{\text{AB}} - \Delta^{\text{Prop}}_{\text{AB}}$) is the more salient gap. Our demonstration of how divergent delivery can lead to magnitude changes and sign reversals, even under relatively easy-to-satisfy conditions, should be of particular interest to strategic marketers, brand managers, and academic researchers.

Advice for Experimenters, Depending On Their Objectives

When experimenters use platforms' A-B testing tools to make comparisons across ads, they may not be learning what they think they are. By considering the distinct objectives that the experimenters may have for their A-B tests, we make prescriptive recommendations to those who are considering using the available ad A-B testing tools. Figure 12 summarizes the decision process an experimenter should go through before deciding if these tools are appropriate for their use case.

For the experimenters whose goal is to predict which ad creatives will "perform best" in a targeted

environment — under the same conditions on the same ad platform with the same campaign settings — our advice is: *Carry on using available* A-B *testing tools*. The current tools are designed to test the bundled value proposition offered by platforms to advertisers: the ability to expose users to ads, intertwined with use of an algorithm to expose the "right" users to the "right" ads. Without needing to separate those two drivers of ads' overall "performance," the experimenters with this goal do not mind—and even may prefer—that their A-B tests lack of balance across ad creative treatments and lack representativeness of the subjects.

Experimenters using these tools for learning about how their different ad creatives generate different responses need to understand how to interpret A-B test results and how to communicate those nuances to managers receiving their analyses. For example, suppose a brand manager were to ask an analytics team to run an A-B test on a predefined audience (such as the high-income homeowners near Houston from our earlier landscaper example). The report of the test should include the disclaimer that the A-B comparisons were made on a subset of the audience, across different mixes of users optimized for each ad separately, where subjects were selected by the proprietary algorithm, based on unknown criteria that cannot be described or enumerated, even by employees of the platform (Gordon et al. 2019). When it comes to affecting the actual A-B testing tools on ad platforms, the least we can do is advocate for transparency and clear language by platforms. That language should more accurately describe what is randomized and what kind of causal inferences can be made.

If the marketing objective is to extrapolate comparisons between ad content for use outside of the current platform (e.g., marketing strategy development, or offline advertising where randomized experimentation and user tracking is more challenging), then our advice is: *Do not rely on these* A-B *tests for causal evidence about the effects of creative content across ads.*. The analytics team, for instance, should warn that results are confounded by how the algorithm determined which ad treatments were most relevant to different experimental subjects. These disclosures should also be made by academic researchers who use A-B test results for scientific inference (Braun et al. 2024).

Limitations and Future Directions

Many of the limitations of our analysis reflect the exact same limitations that experimenters face when using platforms' available A-B testing tools. For instance, we would have liked to have been able to show empirical evidence of just how much estimates of A-B differences vary under different targeting policies, like divergent delivery targeting, proportional targeting, and uniformly random exposure. But not only is it impossible for the experimenter to "turn off" divergent delivery, platforms cannot do it either (see Footnote 21 and Gordon et al. 2023). And even if this kind of study were possible, we still would not be able to describe the extent of divergent delivery beyond the small number of observable aggregate demographic variables in the reported results.

Future work may also consider finding ways to extract information about targeting policies that would allow experimenters to partially adjust aggregated results in some way. Alternatively, there may be specific pieces of information that the platform might be willing to provide that do not infringe on its proprietary interests, yet still quantify or bound the degree to which estimated A-B differences deviate from their hypothetical, balanced baselines.

Finally, we want to point out one reason that disaggregation of A-B test results over a larger set of variables is unlikely. User privacy concerns have motivated recent regulations governing data protection (e.g., GDPR) and security features on certain devices that require websites and apps to get affirmative consent from users before collecting their data. We do not believe that these developments affect the importance of our findings because these mostly deal with *third-party* data, which are collected from sources outside the platform itself through cross-site tracking tools. Platforms will still be able to collect first-party data from their own users (and any second-party data that users might provide to them). While restrictions on cookies and conversion pixels might make holdout tests less effective (or impossible), platforms can still engage in divergent delivery using their own data. As long as platforms continue to consider user-ad relevance when deciding which users see which ads during an A-B test, then all of our concerns about divergent delivery still apply.

References

Ali, Muhammad, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron 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.

Ascarza, Eva (2018). "Retention Futility: Targeting High-Risk Customers Might Be Ineffective." *Journal of Marketing Research*, 60(1):80–98.

Atalay, A. Selin, Siham El Kihal, and Florian Ellsaesser (2023). "Creating Effective Marketing Messages Through Moderately Surprising Syntax." *Journal of Marketing*, 87(5):755–775.

Banker, Sachin and Joowon Park (2020). "Evaluating Prosocial COVID-19 Messaging Frames: Evidence from a Field Study on Facebook." *Judgment and Decision Making*, 15(6):1037–1043.

Barajas, Joel, Ram Akella, Marius Holtan, and Aaron Flores (2016). "Experimental Designs and Estimation for Online Display Advertising Attribution in Marketplaces." *Marketing Science*, 35(3): 465–483.

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

Braun, Michael, Bart De Langhe, Stefano Puntoni, and Eric Schwartz (2024). "Leveraging Digital Advertising Platforms for Consumer Research." *Journal of Consumer Research*, 15(1):119–128.

Cecere, Grazia, Clara Jean, Matthieu Manant, and Catherine Tucker (2018). "Computer Algorithms Prefer Headless Women." *2018 MIT CODE: Conference on Digital Experimentation*.

Cunningham, Scott (2021). *Causal Inference: The Mixtape*. New Haven: Yale University Press.

D.'Angelo, Jennifer K. and Francesca Valsesia (2023). "You Should Try These Together: Combinatory Recommendations Signal Expertise and Improve Product Attitudes." *Journal of Marketing Research*, 60(1):155–169.

De Langhe, Bart and Stefano Puntoni (2021). "Does Personalized Advertising Work as Well as Tech Companies Claim?" *Harvard Business Review*. URL: https://bit.ly/PdeL_HBR.

Eckles, Dean, Brett R. Gordon, and Garrett 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, Brett R., Robert Moakler, and Florian Zettelmeyer (2023). "Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement." *Marketing Science*, 42(4):768–793.

Gordon, Brett R., Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky (2019). "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook." *Marketing Science*, 38(2):193–225.

Hardisty, David J. and Elke U. Weber (2020). "Impatience and Savoring vs Dread: Asymmetries in Anticipation Explain Consumer Time Preferences for Positive vs. Negative Events." *Journal of Consumer Psychology*, 30(4):598–613.

Hutchinson, J. Wesley, Wagner A. Kamakura, and John G. Lynch (2000). "Unobserved Heterogeneity as an Alternative Explanation for Reversal Effects in Behavioral Research." *Journal of Consumer Research*, 27(3):324–344.

Johnson, Garrett A. (2023). "Inferno: A Guide to Field Experiments in Online Display Advertising." *Journal of Economics and Management Strategy*.

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

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

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

Lambrecht, Anja and Catherine Tucker (2019). "Algorithmic Bias? An Empirical Study of Apparent Gender Based Discrimination in the Display of STEM Career Ads." *Management Science*, 65(7): 2966–2981.

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

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.

Mookerjee, Siddhanth, Yann Cornil, and JoAndrea Hoegg (2021). "From Waste to Taste: How 'Ugly' Labels Can Increase Purchase of Unattractive Produce." *Journal of Marketing*, 85(3):62–77.

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

Paharia, Neeru (2020). "Who Receives Credit or Blame? The Effects of Made-to-Order Production on Responses to Unethical and Ethical Company Production Practices." *Journal of Marketing*, 84(1): 88–104.

Paharia, Neeru and Vanitha Swaminathan (2019). "Who Is Wary of User Design? The Role of Power-Distance Beliefs in Preference for User-Designed Products." *Journal of Marketing*, 83(3):91–107.

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

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

Winterich, Karen Page, Gergana Y. Nenkov, and Gabriel E. Gonzales (2019). "Knowing What It Makes: How Product Transformation Salience Increases Recycling." *Journal of Marketing*, 83(4): 21–37.

Zhou, Lingrui, Katherine M. Du, and Keisha M. Cutright (2022). "Befriending the Enemy: The Effects of Observing Brand-to-Brand Praise on Consumer Evaluations and Choices." *Journal of Marketing*, 86(4):57–72.

Web Appendices

Where A-B Testing Goes Wrong: How Divergent Delivery Affects What Online Experiments Cannot (and Can) Tell You About How Customers Respond to Advertising

Michael Braun and Eric Schwartz

July 30, 2024

Contents

WEB APPENDIX A: DETROIT EMPLOYMENT MARKETING EXPERIMENT

Remaining Ad Treatments

These images are the remaining eight figures used in the Detroit recruiting example (the other six are in Figure 4 in the main text).

Crime Analyst, Self, and Rational

Crime Analyst, Other, and Emotional

Dom Violence Unit, Other, and Emotional

Dom Violence Unit, Other, Rational

Dom Violence Unit, Self, Rational

Detroit Police Department
Sponsored ·

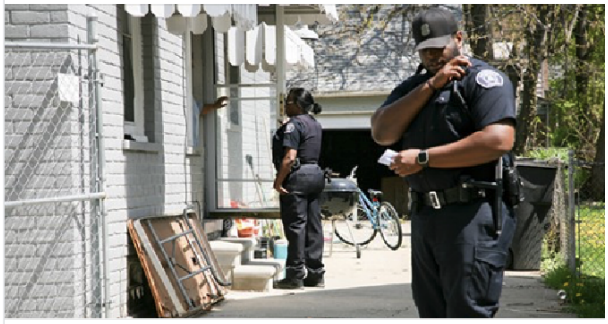
Domestic violence may happen. Thanks to the caring officers in the Domestic Violence Unit and others at DPD, these crimes get solved. DPD offers training and job opportunities in many career paths.

WWW.DETROITMI.GOV/JOINDPD
Advance your career and develop your skills.
We are hiring! The Police Academy is now accepting…
Learn More

Patrol Officer, Other, and Emotional

Detroit Police Department
Sponsored ·

Maya was a victim of robbery. Thanks to the vigilant patrol officers and others at DPD, she now feels safer in her neighborhood. DPD offers opportunities to make a difference by helping the people of Detroit.

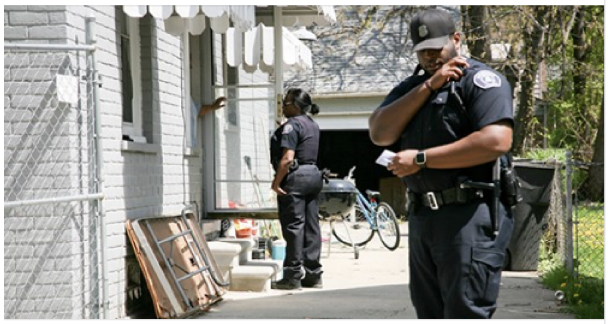


WWW.DETROITMI.GOV/JOINDPD
Serve your community and keep Detroit safe.
We are hiring! The Police Academy is now accepting…
Learn More

Patrol Officer, Self, and Rational

Detroit Police Department
Sponsored ·

Robberies happen. Thanks to the vigilant patrol officers and others at DPD, these crimes get solved. DPD offers training and job opportunities in many career paths.

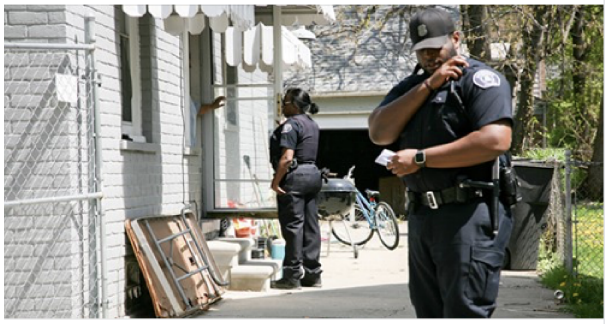


WWW.DETROITMI.GOV/JOINDPD
Advance your career and develop your skills.
We are hiring! The Police Academy is now accepting…
Learn More

Patrol Officer, Self, and Emotional

Detroit Police Department
Sponsored ·

Maya was a victim of robbery. Thanks to the vigilant patrol officers and others at DPD, she now feels safer in her neighborhood. DPD offers training and job opportunities in many career paths.



WWW.DETROITMI.GOV/JOINDPD
Advance your career and develop your skills.
We are hiring! The Police Academy is now accepting…
Learn More

Data

Table WA1: Impressions and Unique Users Targeted Per Ad Treatment

Ad Creatives			Impressions		Unique Users	
Job	Rat/Emot	Other/Self	Total	Prop. Female	Total	Prop. Female
Data Analyst	Rational	Other	38,617	.549	6,724	.484
		Self	40,778	.541	6,646	.494
	Emotional	Other	36,434	.557	6,644	.504
		Self	38,548	.554	6,826	.508
Dom. Violence	Rational	Other	36,853	.590	6,856	.529
		Self	40,064	.560	7,378	.497
	Emotional	Other	35,695	.588	7,036	.534
		Self	37,538	.592	6,980	.530
Patrol	Rational	Other	35,604	.487	7,200	.484
		Self	37,651	.476	7,710	.464
	Emotional	Other	38,190	.487	7,586	.473
		Self	36,651	.482	7,390	.474
Control		ESDC	39,037	.659	5,362	.570
		PAL	41,501	.546	5,812	.508

NOTE: Prop. Female is the proportion of unique users whom Facebook categorized as female. The Control ad creatives were for nonprofit organizations supporting children, Every School Day Counts and Police Athletic League. The experiment was conducted using the Facebook Marketing API, version 3.3.

WEB APPENDIX B: DEFINING THE CONFOUND UNDER DIVERGENT DELIVERY

As defined in Equation 2, the change in incremental lift caused by exposure to ad A relative to B is

$$\Delta^{i}_{AB} = \left(Y^{(1)}_{i,A} - Y^{(0)}_{i,A}\right) - \left(Y^{(1)}_{i,B} - Y^{(0)}_{i,B}\right). \tag{B1}$$

From Table 2, the expected A-B difference, conditional on user type, is

$$\Delta^{X}_{AB} = \mathbf{E}\Big[\left(Y^{(1)}_{i,A} - Y^{(0)}_{i,A}\right) - \left(Y^{(1)}_{i,B} - Y^{(0)}_{i,B}\right) \,|\, X_i = X\Big] \tag{B2}$$

This expression does not depend on Z_i because all users are endowed with a complete set of potential outcomes for all ads. Specifically, every user has all four potential outcomes, in the case of an A-B test. That is, ad assignment Z_i affects which of the potential outcomes is realized, but not the means or whole distributions of the potential outcomes themselves.

The posterior type distributions for ads A and B are defined in Table 1 as

$$\Pi^{Targ}_{X|A} = \mathbf{P}\Big(X_i = X \,|\, \tau^{i}_{A} = 1, Z_i = A\Big) \tag{B3}$$

$$\Pi^{Targ}_{X|B} = \mathbf{P}\Big(X_i = X \,|\, \tau^{i}_{B} = 1, Z_i = B\Big) \tag{B4}$$

The following lifts are mixtures of λ_{XZ} over $\Pi^{Targ}_{X|Z'}$, where Z' may or may not equal Z. The effect of ad A could be defined with respect to users eligible for and targeted with *either* ad A or B, and vice-versa. For instance, we consider the expected potential outcomes of users responding to B among users actually assigned to ad A. Here are all four possibilities:

$$\lambda^{Targ_A}_{A} = \mathbf{E}\Big[Y^{(1)}_{i,A} - Y^{(0)}_{i,A} \,|\, \tau^{i}_{A} = 1, Z_i = A\Big] \tag{B5}$$

$$\lambda^{Targ_B}_{A} = \mathbf{E}\Big[Y^{(1)}_{i,A} - Y^{(0)}_{i,A} \,|\, \tau^{i}_{B} = 1, Z_i = B\Big] \tag{B6}$$

$$\lambda^{Targ_A}_{B} = \mathbf{E}\Big[Y^{(1)}_{i,B} - Y^{(0)}_{i,B} \,|\, \tau^{i}_{A} = 1, Z_i = A\Big] \tag{B7}$$

$$\lambda^{Targ_B}_{B} = \mathbf{E}\Big[Y^{(1)}_{i,B} - Y^{(0)}_{i,B} \,|\, \tau^{i}_{B} = 1, Z_i = B\Big] \tag{B8}$$

Note that $\lambda^{Targ}_{A} = \lambda^{Targ_A}_{A}$ and $\lambda^{Targ}_{B} = \lambda^{Targ_B}_{B}$. But $\lambda^{Targ_B}_{A} \neq \lambda^{Targ}_{A}$ and $\lambda^{Targ_A}_{B} \neq \lambda^{Targ}_{B}$ *unless* $\Pi^{Targ}_{X|A} = \Pi^{Targ}_{X|B}$, meaning the mixes targeted with each ad are the same.

It is not straightforward to define a theoretical quantity Δ^{Targ}_{AB} as an A-B difference in potential outcomes over a single group of "targeted users" since the mix of targeted users for ad A differs from that of ad B. Therefore, we separately define $\Delta^{Targ_A}_{AB}$ for the mix targeted with A and $\Delta^{Targ_B}_{AB}$ for the mix targeted with B. These two are mixtures of Δ^{X}_{AB} over $\Pi^{Targ}_{X|A}$ and $\Pi^{Targ}_{X|B}$, respectively.

$$\Delta_{AB}^{Targ_A} = \sum_{\forall X} \Delta_{AB}^{X} \Pi_{X|A}^{Targ} \tag{B9}$$

$$= \mathbf{E}\Big[\left(Y_{i,A}^{(1)} - Y_{i,A}^{(0)} \right) - \left(Y_{i,B}^{(1)} - Y_{i,B}^{(0)} \right) \mid \tau_A^i = 1, Z_i = A \Big] \tag{B10}$$

$$= \mathbf{E}\Big[Y_{i,A}^{(1)} - Y_{i,A}^{(0)} \mid \tau_A^i = 1, Z_i = A \Big] - \mathbf{E}\Big[Y_{i,B}^{(1)} - Y_{i,B}^{(0)} \mid \tau_A^i = 1, Z_i = A \Big] \tag{B11}$$

$$= \lambda_A^{Targ} - \lambda_B^{Targ_A} \tag{B12}$$

$$\Delta_{AB}^{Targ_B} = \sum_{\forall X} \Delta_{AB}^{X} \Pi_{X|B}^{Targ} \tag{B13}$$

$$= \mathbf{E}\Big[\left(Y_{i,A}^{(1)} - Y_{i,A}^{(0)} \right) - \left(Y_{i,B}^{(1)} - Y_{i,B}^{(0)} \right) \mid \tau_B^i = 1, Z_i = B \Big] \tag{B14}$$

$$= \mathbf{E}\Big[Y_{i,A}^{(1)} - Y_{i,A}^{(0)} \mid \tau_B^i = 1, Z_i = B \Big] - \mathbf{E}\Big[Y_{i,B}^{(1)} - Y_{i,B}^{(0)} \mid \tau_B^i = 1, Z_i = B \Big] \tag{B15}$$

$$= \lambda_A^{Targ_B} - \lambda_B^{Targ} \tag{B16}$$

If $\Pi_{X|A}^{Targ} = \Pi_{X|B}^{Targ}$, then

$$\sum_{\forall X} \lambda_{XA} \Pi_{X|A}^{Targ} = \sum_{\forall X} \lambda_{XA} \Pi_{X|B}^{Targ} \tag{B17}$$

$$\mathbf{E}\Big[Y_{i,A}^{(1)} - Y_{i,A}^{(0)} \mid \tau_A^i = 1, Z_i = A \Big] = \mathbf{E}\Big[Y_{i,A}^{(1)} - Y_{i,A}^{(0)} \mid \tau_B^i = 1, Z_i = B \Big] \tag{B18}$$

$$\lambda_A^{Targ} = \lambda_A^{Targ_B} \tag{B19}$$

and

$$\sum_{\forall X} \lambda_{XB} \Pi_{X|A}^{Targ} = \sum_{\forall X} \lambda_{XB} \Pi_{X|B}^{Targ} \tag{B20}$$

$$\mathbf{E}\Big[Y_{i,B}^{(1)} - Y_{i,B}^{(0)} \mid \tau_A^i = 1, Z_i = A \Big] = \mathbf{E}\Big[Y_{i,B}^{(1)} - Y_{i,B}^{(0)} \mid \tau_B^i = 1, Z_i = B \Big] \tag{B21}$$

$$\lambda_B^{Targ} = \lambda_B^{Targ_B} \tag{B22}$$

So in the case of $\Pi_{X|A}^{Targ} = \Pi_{X|B}^{Targ}$, we have $\Delta_{AB}^{Targ_A} = \Delta_{AB}^{Targ_B} = \lambda_A^{Targ} - \lambda_B^{Targ}$. Since this condition occurs under proportional targeting, we call this common A-B difference Δ_{AB}^{Prop}. This A-B difference has internal validity because A and B are compared against the same mix of user types. Its values do not depend on whether targeted users were eligible to receive A or B.

However, under divergent delivery, $\Pi_{X|A}^{Targ} \neq \Pi_{X|B}^{Targ}$, which means that $\Delta_{AB}^{Targ_A} \neq \Delta_{AB}^{Targ_B}$. That means that $\lambda_A^{Targ} - \lambda_B^{Targ}$ equals some value that is different from Δ_{AB}^{Prop}. The estimated difference, $\widehat{\Delta}_{AB}^{Targ} = \widehat{\lambda}_A^{Targ} - \widehat{\lambda}_B^{Targ}$ contains a confound because it is "contaminated" by the variation in the type distributions across users eligible for the two different ads.

WEB APPENDIX C: SIMULATION DETAILS

The simulation has four stages:

- Define 800 audience "worlds," each with a distinct set of parameters for targeting policies and expected potential outcomes.
- Simulate 15 audiences within each world, each with $N = 200{,}000$ users, half with $X_i = P$ and half with $X_i = Q$.
- For each user, generate user-level potential outcomes.
- For each user, simulate targeting decisions and lift study arm assignments (which determine which potential outcomes are realized).

In addition to symbols defined in the main text, we use the following symbols in our description of the simulation.

$$\alpha_\tau = \Phi_{\bullet A} / \Phi_{\bullet B} \qquad \alpha_Y = \mu_A^{(1)} / \mu_B^{(1)} \tag{C1}$$

$$\pi_\tau = \Phi_{P\bullet} / \Phi_{Q\bullet} \qquad \pi_Y = \mu_P^{(1)} / \mu_Q^{(1)} \tag{C2}$$

$$\Phi_{\bullet\bullet} = \sum_{\forall Z} \Phi_{\bullet Z} \mathbf{P}(Z_i = Z) \quad \text{(aggregate probability that any user is targeted)} \tag{C3}$$

$$R_i = \begin{cases} 1 & \text{if user i is assigned to the treatment arm of a lift study} \\ 0 & \text{if user i is assigned to the holdout arm of a lift study} \end{cases} \tag{C4}$$

A "world" consists of a complete set of parameters for user response parameters and targeting policies.

1. Set parameters for 800 "audience worlds," which are defined in Table WC1.
2. Transform the parameters for each world into targeting probabilities and expected potential outcomes.

   (a) Set the following intermediate values.

$$S_\tau \leftarrow \sqrt{(\alpha_\tau \pi_\tau - 1)^2 + (\alpha_\tau - \pi_\tau)^2 \rho_\tau^2 + 2\rho_\tau (\alpha_\tau \pi_\tau (\alpha_\tau + \pi_\tau + 4) + \alpha_\tau + \pi_\tau)} \tag{C5}$$

$$S_Y \leftarrow \sqrt{(\alpha_Y \pi_Y - 1)^2 + (\alpha_Y - \pi_Y)^2 \rho_Y^2 + 2\rho_Y (\alpha_Y \pi_Y (\alpha_Y + \pi_Y + 4) + \alpha_Y + \pi_Y)} \tag{C6}$$

$$F_\tau \leftarrow (\alpha_\tau + 1)(\pi_\tau + 1)(\rho_\tau - 1) \tag{C7}$$

$$F_Y \leftarrow (\alpha_Y + 1)(\pi_Y + 1)(\rho_Y - 1) \tag{C8}$$

Table WC1: Simulation Parameters

Common for all worlds		
$\widetilde{\mu}^{(1)}$	$= .2$	Aggregate conversion rate
α_Y	$= 1.15$	Ratio of conversion rates of A to B
$\Phi_{\bullet\bullet}$	$= .2$	Aggregate targeting probability
α_τ	$= 1$	Ratio of targeting proportions of A to B
Discrete experimental conditions (20 total)		
π_Y	$\in \{1, 1/3\}$	Ratio of aggregate conversion rates of Poets and Quants
ρ_Y	$\in \{1, 4\}$	Either proportional or divergent response
ρ_τ	$\in \{1/8, 1/4, 1, 4, 8\}$	Levels of divergent delivery
Simulated parameters (40 for each discrete parameter)		
$\mu_P^{(0)}$	$\sim$ Unif(.02,.04)	Expected conversion rates for Poets who are not exposed to an ad.
$\mu_Q^{(0)}$	$\sim$ Unif(.02,.04)	Expected conversion rates for Quants who are not exposed to an ad.
π_Y	$\sim \log_2 \text{Unif}(1/5, 5)$[2]	Ratio of targeting proportions of Poets to Quants.

NOTE: To clarify, there are 40 audiences with the same values of π_Y, ρ_Y, and ρ_τ. Each of those audiences has a different simulated value for π_τ, $\mu_P^{(0)}$, and $\mu_Q^{(0)}$.

(b) Construct the targeting policy for the audience.[1]

$$\Phi_{PA} \leftarrow \frac{2\Phi_{\bullet\bullet}}{F_\tau}\left(\rho_\tau(\alpha_\tau + \pi_\tau + 2\alpha_\tau\pi_\tau) - \alpha_\tau\pi_\tau - S_\tau + 1\right) \tag{C9}$$

$$\Phi_{PB} \leftarrow \frac{2\Phi_{\bullet\bullet}}{F_\tau}\left(\pi_\tau(\rho_\tau - 2) - \alpha_\tau(\pi_\tau + \rho_\tau) + S_\tau - 1\right) \tag{C10}$$

$$\Phi_{QA} \leftarrow \frac{2\Phi_{\bullet\bullet}}{F_\tau}\left(\alpha_\tau(\rho_\tau - 2) - \pi_\tau(\alpha_\tau + \rho_\tau) + S_\tau - 1\right) \tag{C11}$$

$$\Phi_{QB} \leftarrow \frac{2\Phi_{\bullet\bullet}}{F_\tau}\left(\rho_\tau(\alpha_\tau + \pi_\tau + 2) + \alpha_\tau\pi_\tau - S_\tau - 1\right) \tag{C12}$$

(c) Set the remaining expected potential outcomes.

$$\mu_{PA}^{(1)} \leftarrow \frac{2\widetilde{\mu}^{(1)}}{F_Y}\left(\rho_Y(\alpha_Y + \pi_Y + 2\alpha_Y\pi_Y) - \alpha_Y\pi_Y - S_Y + 1\right) \tag{C13}$$

$$\mu_{PB}^{(1)} \leftarrow \frac{2\widetilde{\mu}^{(1)}}{F_Y}\left(\pi_Y(\rho_Y - 2) - \alpha_Y(\pi_Y + \rho_Y) + S_Y - 1\right) \tag{C14}$$

$$\mu_{QA}^{(1)} \leftarrow \frac{2\widetilde{\mu}^{(1)}}{F_Y}\left(\alpha_Y(\rho_Y - 2) - \pi_Y(\alpha_Y + \rho_Y) + S_Y - 1\right) \tag{C15}$$

$$\mu_{QB}^{(1)} \leftarrow \frac{2\widetilde{\mu}^{(1)}}{F_Y}\left(\rho_Y(\alpha_Y + \pi_Y + 2) + \alpha_Y\pi_Y - S_Y - 1\right) \tag{C16}$$

For each audience, simulate experimental results based on parameters of that audience's world.

3. For all users $i = 1, \ldots, N$, sample potential outcomes $Y_{i,A}^{(1)} \sim \text{Bernoulli}\left(\mu_{XA}^{(1)}\right)$, $Y_{i,B}^{(1)} \sim \text{Bernoulli}\left(\mu_{XB}^{(1)}\right)$, and $Y_i^{(0)} \sim \text{Bernoulli}\left(\mu_X^{(0)}\right)$.

[1] In Steps 2b and 2c, dividing by F_τ and F_Y from Step 2a creates removable discontinuities at $\rho_\tau = 1$ and $\rho_Y = 1$. Adding a small value like 10^{-10} to ρ_τ and ρ_Y is a sufficient remedy.

4. For $Z \in \{A,B\}$, compute $\mu^{(1)}_{Z,\text{Aud}}$ and $\mu^{(0)}_{Z,\text{Aud}}$ as averages of potential outcomes.

5. Determine each user's *realized* outcome by simulating the eligibility, targeting, and holdout processes.

   (a) Assign eligible ads to users: sample Z_i from $Z \in \{A,B\}$ with $\mathbf{P}(Z_i = A) = .5$.

   (b) Target users conditional on user type and assigned ad: $\tau^i_{Z_i} \sim \text{Bernoulli}\left(\Phi_{X_i Z_i}\right)$.

   (c) Assign targeted users to arms of the holdout test: $R_i \sim \text{Bernoulli}(.5)$.

6. Compute observable statistics for each audience by averaging over realized outcomes.

   (a) For $Z \in \{A,B\}$, compute $\bar{Y}^{(1)}_{Z,\text{Targ}}$ as an average of $Y^{(1)}_{i,Z}$ among users with $Z_i = Z$, $\tau^i_Z = 1$, and $R_i = 1$.

   (b) For $Z \in \{A,B\}$, compute $\bar{Y}^{(0)}_{Z,\text{Targ}}$ as an average of $Y^{(0)}_{i,Z}$ among users with $Z_i = Z$, $\tau^i_Z = 1$, and $R_i = 0$.

   (c) Compute λ^{Aud}_A, λ^{Aud}_B, Δ^{Aud}_{AB}, $\widehat{\lambda}^{\text{Targ}}_A$, $\widehat{\lambda}^{\text{Targ}}_B$, and $\widehat{\Delta}^{\text{Targ}}_{AB}$ using definitions in Table 2.

   (d) λ^{Prop}_A, λ^{Prop}_B, $\Delta^{\text{Prop}}_{AB}$ are the values of $\widehat{\lambda}^{\text{Targ}}_A$, $\widehat{\lambda}^{\text{Targ}}_B$, and $\widehat{\Delta}^{\text{Targ}}_{AB}$ computed from the $\rho_\tau = 1$ conditions.