GARRETT A. JOHNSON, RANDALL A. LEWIS, and ELMAR I. NUBBEMEYER*

To measure the effects of advertising, marketers must know how consumers would behave had they not seen the ads. The authors develop a methodology they call "ghost ads," which facilitates this comparison by identifying the control group counterparts of the exposed consumers in a randomized experiment. The authors show that, relative to public service announcement and intent-to-treat A/B tests, ghost ads can reduce the cost of experimentation, improve measurement precision, deliver the relevant strategic baseline, and work with modern ad platforms that optimize ad delivery in real time. The authors also describe a variant, "predicted ghost ad" methodology, which is compatible with online display advertising platforms; their implementation records more than 100 million predicted ghost ads per day. The authors demonstrate the methodology with an online retailer's display retargeting campaign. They show novel evidence that retargeting can work: the ads lifted website visits by 17.2% and purchases by 10.5%. Compared with intent-to-treat and public service announcement experiments, advertisers can measure ad lift just as precisely while spending at least an order of magnitude less.

Keywords: field experiments, advertising effectiveness, digital advertising

Online Supplement: http://dx.doi.org/10.1509/jmr.15.0297

Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness

Marketers need to know the effectiveness of their advertising. Digitization has advanced this goal by providing marketers unprecedented access to data on clicks, site visits, and online purchases by individual consumers who see their ads. While granular measurement is now routine in digital advertising, rigorous causal measurement of ad effectiveness is not (Lavrakas 2010). Randomized controlled experiments can be a simple and effective tool for causal measurement, but current approaches limit their widespread use.

*Garrett Johnson is Visiting Assistant Professor of Marketing, Kellogg School of Management, Northwestern University (email: garrett.johnson@kellogg.northwestern.edu). Randall Lewis is Economic Research Scientist, Netflix (email: randall@econinformatics.com). Elmar Nubbemeyer is Product Manager, Netflix (email: elmar.nubbemeyer@gmail.com). The authors thank seminar participants at Columbia University, Kellogg School of Management, Stanford Graduate School of Business, and Rady School of Management, as well as Abdelhamid Abdou, Eric Anderson, David Broockman, Hubert Chen, Brett Gordon, Mitch Lovett, Preston McAfee, John Pau, David Reiley, Stephan Seiler, Robert Saliba, Kathryn Shih, Robert Snedegar, Hal Varian, Ken Wilbur, three anonymous referees, and many Google employees and advertisers for contributing to the success of this project. The second and third authors acknowledge that they were employed by Google and owned Google stock while conducting this research. Coeditor: Rajdeep Grewal; Associate Editor: Peter Danaher.

A popular approach for advertising effectiveness experiments is delivering public service announcements (PSAs) to the control group. PSAs identify baseline purchase behavior among the subset of control consumers reached by the ads: the experimental ad effect estimator compares outcomes between people who see the focal advertiser's ad and those who see the PSA. Unfortunately, PSAs are expensive and prone to error because they require coordination among advertisers, publishers, and third-party charities. These costs reduce advertisers' incentives to learn through experimentation (Gluck 2011).

Worse, PSA experiments are rendered invalid when marketers use computer algorithms to optimize ad delivery in-campaign separately for the PSA and focal campaigns. These algorithms use machine learning models to maximize consumer clicks, site visits, or purchases. In-campaign optimization breaks PSA experiments: to maximize performance of the focal ad and PSA, the ad platform will assign different types of consumers to be exposed to the PSA or the focal ad. However, an experiment is predicated on the symmetry of the treatment and control groups: the experimental groups must be the same except for the focal ad. Thus, the PSA-exposed consumers are no longer a valid holdout group for the focal ad–exposed consumers. This failure of PSAs is worrisome for

ISSN: 0022-2437 (print)
1547-7193 (electronic)

Journal of Marketing Research
Vol. LIV (December 2017), 867–884
DOI: 10.1509/jmr.15.0297

the online display ad industry, considering that two-thirds of its spending uses automated in-campaign optimization (Interactive Advertising Bureau 2014).

We propose a new methodology for ad effectiveness field experiments: ghost ads. Like PSAs, ghost ads identify ads in the control group that *would have been* the focal advertiser's ads had the consumer been in the treatment group; as such, ghost ads deliver valid estimates of the average treatment effect on the treated. Unlike PSAs, ghost ads are compatible with in-campaign optimization technology and avoid the costs of PSAs. Instead of PSA ads, the control group consumer sees whatever ads the ad platform chooses to deliver when the focal ad is absent. The ghost ad methodology is implemented at the ad platform level. Typically, the ad platform runs an auction to determine which ad it will show to a consumer and then records these ads in a database. To determine when the focal ad would have been shown to a consumer in the control group, the ad platform can run a second, simulated auction that includes the focal ad in the set of potential ads. The ad platform then records *ghost ad impressions*—the would-be focal ad impressions in the control group—in a second database. Ghost ads are so named because they make the experimental ads visible to the ad platform and experimenter, but invisible to the control group consumers. The ghost ad method can be applied across several digital media—for example, search and online display (including audio and video)—whenever consumers can be randomly sorted into treatment groups.

One contribution of the ghost ad methodology is that it allows researchers to design valid experiments that leverage the algorithmic nature of the exposure decision. Incomplete exposure or treatment is a familiar problem in the experimental literature, though the traditional focus has been on the subject's compliance with treatment. Our setting is novel because the focal advertiser and the ad platform partially abdicate the ad exposure decision to an algorithm. Our insight is to leverage the same algorithm to replicate the exposure decision in the control group. Because ad platforms are not designed to optimize campaigns on the basis of another campaign's performance, we show that ghost ads can be implemented using a simulated auction.

Ghost ads also deliver the relevant baseline behavior in the control group that reflects the focal advertiser's strategic environment. The focal advertiser competes with other advertisers in a marketplace for consumers' attention. When the focal advertiser exits the marketplace, many other advertisers take its place, including some direct competitors—which creates an externality on the focal advertiser. For instance, advertising can play a "defensive" role in the sense that it blocks competitors from stealing consumers, in addition to its "offensive" role of pulling in consumers. Because PSAs are chosen to be orthogonal to the advertiser, estimates using them ignore the strategic consequences of the competing ads they displace. If advertising has a defensive role, PSAs will understate the total effect of the ads relative to ghost ad estimates.

We propose a second, more robust methodology we call "predicted ghost ads" to be used when the ad platform and a downstream firm jointly control ad delivery. For instance, many online display ad platforms only suggest an ad to a downstream web publisher, which can reject the ad for many reasons. A new approach is needed because the naive ghost ad approach cannot observe control group cases in which the platform suggests the focal ad but the publisher refuses it. This slippage breaks the symmetry between the focal ads in the treatment group and the ghost ads in the control group. To offset this shortcoming, the predicted ghost ad methodology preserves this symmetry by running a simulated auction across *both* treatment and control group consumers to determine whether the platform intends to deliver an experimental ad before determining the real winner in both cases. The ad platform then logs predicted ghost ad impressions in another database to flag occasions when the ad platform intends to serve an experimental ad regardless of the consumer's actual treatment assignment. Researchers can use the predicted ghost ad impressions to construct a powerful instrumental variable to calculate the local average treatment effect (LATE; see Imbens and Angrist 1994) among consumers who are predicted to be exposed to the focal ad. We implement predicted ghost ads on Google's online display network and demonstrate it using an application. Currently, predicted ghost ads are broadly used, generating more than 100 million predicted impressions daily. In follow-up work, we describe the results of more than 400 advertiser field experiments averaging 4 million users each (Johnson, Lewis, and Nubbemeyer 2017).

We apply our predicted ghost ad methodology to show novel evidence that an online display retargeting campaign can generate incremental conversions. The effectiveness of retargeting ads is controversial. Although consumers who see retargeted ads may have high sales, this association may not be causal because the exposed consumers are a highly self-selected group that may have purchased anyway. Retargeting could even reduce ad effectiveness if it provokes reactance in consumers. Whereas Lambrecht and Tucker (2013) and Bleier and Eisenbeiss (2015) compare the relative performance of different retargeting creatives, we provide the first evidence that retargeting can work when compared with a control. Working with an online sports and outdoors retailer to retarget consumers who visited product pages on its website, we find that the retargeting campaign increases website visits by 17.2% ($t = 13.64$), transactions by 11.9% ($t = 5.54$), and sales by 10.5% ($t = 3.51$). This retargeting campaign was performance optimized and incompatible with PSAs, but our predicted ghost ad methodology succeeds and demonstrates highly significant lifts.

The article is organized as follows. The next section describes related literature. The following section outlines the ad effectiveness measurement problem and explains the existing PSA and intent-to-treat (ITT) approaches. We then explain the ghost ad methodology and describe its benefits. The following section describes two related methodologies: predicted ghost ads and ghost bids. Next, we describe our empirical application and provide evidence that a retargeting campaign can increase sales. The final two sections outline some challenges in implementing ghost ads and suggest future uses for the technique.

LITERATURE REVIEW

The high cost of running ad experiments limits their use and affects the form they take. Ghost ads are designed to eliminate an important cost of experiments—the cost of PSAs—because this cost discourages advertisers from using PSA tests. For example, of the 25 online display experiments at Yahoo listed by Lewis and Rao (2015), only 8 use PSAs, and the average control group allocation is only a third of users, even though an even split would maximize precision. When advertisers pay for PSAs, they typically prefer to reduce costs by making the control groups small (sometimes as small as 2%; Hoban and Bucklin 2015), which sacrifices measurement precision.

In addition, PSA tests—especially those with even treatment/control splits—often arise out of exceptional circumstances. First, the ad platform may subsidize an experiment to learn or promote its effectiveness, as in Sahni (2015) or Johnson, Lewis, and Reiley (2017). Second, advertisers may engage research firms to design the experiment as in Goldfarb and Tucker (2011) and Bart, Stephen, and Sarvary's (2014) brand survey studies. Third, Lewis (2014) and Lewis and Nguyen (2015) use natural experiments in which the ad platform randomly splits the delivery of unrelated ad campaigns.

Although ITT A/B[1] experiments also eliminate the costs of PSAs, ITT experiments lack the measurement precision of ghost ads or PSAs. Without PSAs, we cannot observe the control group consumers who would be exposed to an ad when the campaign's reach is incomplete. As the subsection "ITT Approach" elaborates, ITT experiments compare all treatment- and control-eligible consumers—including unexposed consumers—who contribute only noise to the estimator. As such, this approach requires treatments with high advertising expenditure to detect significant effects. Examples include the online display ad experiments by Gordon et al. (2017) at Facebook and Lewis and Reiley (2014) at Yahoo as well as the search ad experiments by Blake, Nosko, and Tadelis (2015) and Kalyanam et al. (2015). As in those studies, ITT is compatible with geographic- rather than consumer-level randomization when the latter is infeasible.

To avoid the cost of PSAs and the opportunity cost of not advertising to a holdout group, many advertisers prefer to measure the relative effectiveness of ads using weight tests, which vary the quantity of ads, or copy tests, which vary creative content. Most of the split-cable TV advertising studies in Lodish et al. (1995) and Hu, Lodish, and Krieger (2007) are either copy tests or weight tests rather than PSA tests. Simester et al. (2009) use a weight test to measure catalog effectiveness. Copy tests compare the performance of different ad creatives, for instance, to assess the effect of social cues (Bakshy et al. 2012) or native advertising (Sahni and Nair 2016). Nonetheless, both copy tests and PSA tests are biased when they employ performance-optimizing technology.

Past retargeting studies use copy tests to evaluate the personalization of retargeting creatives. Lambrecht and Tucker (2013) compare generic brand creatives with dynamic retargeting creatives that include the product with which the consumer engaged (the "focal product"). Bleier and Eisenbeiss (2015) compare retargeting ads that feature products from the focal product's category and/or manufacturer's brand to randomly selected products. Lambrecht and Tucker (2013) find that personalized ads reduce transactions overall, whereas Bleier and Eisenbeiss (2015) find that click-through rates increase. Both studies agree that content targeting is important for personalized retargeting ads whose effectiveness improves when the consumer browses content related to the focal product (Lambrecht and Tucker 2013) or browses shopping sites (Bleier and Eisenbeiss 2015). In contrast, we evaluate the effectiveness of retargeting against a holdout group and include revenues as an outcome. This method allows us to determine whether retargeting causes incremental purchases or merely reaches consumers who would purchase anyway.

Recently, two other retargeting field experiments using holdouts confirm that the strategy can be effective. Sahni, Narayanan, and Kalyanam (2016) show that retargeting can increase site visits even in a high-cost and high-involvement product category. Using a design that varies the timing of the retargeting campaign, the authors show that retargeting is more effective in the days following the user's site visit. Moriguchi, Xiong, and Luo (2016) show that email retargeting can be an effective means for a retailer to increase purchases among online cart abandoners.

The challenge of low statistical power in the ad effectiveness setting must be met with the most precise estimation method available. As Lewis and Rao (2015) explain, the effects of ads are so small relative to the volatility of sales data that informative experiments may require more than 10 million consumer-week observations. Thus, more efficient measurement can make experiments accessible to more advertisers. Lewis, Rao, and Reiley (2015) mention the concept of ghost ads and foreshadow its importance for large-scale experimentation. In this paper, we explicate the ghost ad methodology, its implementation, and its advantages over existing methodologies.

EXISTING EXPERIMENTAL APPROACHES

In this section, we introduce the marketer's problem of measuring ad effectiveness and describe two existing experimental approaches to that problem. This section's exposition borrows from related discussions in Manski (2007) and Imbens and Rubin (2015).

Measuring Ad Effectiveness

Marketers want to know the effectiveness of their advertising. To do so, they should compare their marketing outcomes when they advertise with the outcomes that would have transpired had they not advertised.

Formally, let $D_i \in \{0, 1\}$ be the active treatment received by consumer $i \in \{1, ..., N\}$, where i is either exposed to at least one of the focal advertiser's ads ($D_i = 1$) or not ($D_i = 0$). Let $[Y_i(1), Y_i(0)]$ denote consumer i's potential outcomes as a function of i's active treatment. For each consumer, we observe only the outcome of either being exposed, $Y_i(1)$, or not, $Y_i(0)$, but not both. We denote i's realized outcome by $Y_i \equiv Y_i(D_i)$. The marketer then defines average ad effectiveness among exposed consumers as the average treatment effect on the treated (ATET): $E[Y(1) - Y(0)|D = 1]$.

Measuring ATET is challenging in part because exposure D is not randomly assigned to consumers. Rather, marketers target segments of consumers, and today's advanced ad platforms can further refine these segments to maximize the ad campaign's performance. In addition, consumers choose how much media they consume and therefore determine their number of opportunities to receive the focal ad. We can resolve these problems in the context of a randomized experiment. Let $Z_i \in \{T, C\}$ denote i's treatment assignment to either the treatment group T or control group C. Then, i's active treatment also has $[D_i(T), D_i(C)]$ as potential outcomes that are a function of i's treatment assignment. If assigned to the treatment group, i may or may not be exposed to the focal ad: $D_i(T) \in \{0, 1\}$. If

[1]ITT experiments are similar to A/B or bucket tests in that A/B testing platforms typically use the ITT estimator when analyzing experiments. However, we distinguish ITT experiments from classical A/B tests to emphasize that ad exposure is endogenous, exhibiting partial compliance in our setting; in other words, not all consumers are exposed (see the "Existing Experimental Approaches" section).

assigned to the control group, i is held out from receiving the focal ad: $D_i(C) = 0$.

The ad experiment rests on some basic assumptions. We assume that treatment assignment can only affect outcomes through the active treatment (the exclusion restriction), so that we can write $Y_i(D_i)$ as a function of the active treatment alone. We assume that Z_i is assigned randomly so that Z_i is statistically independent of potential outcomes $[Y_i(1), Y_i(0)]$ and $[D_i(T), D_i(C)]$. We implicitly assume that $Y_i(D_i)$ does not depend on the treatment to which other consumers are assigned and that there are no different forms of each treatment with different potential outcomes—the stable unit treatment value assumption (see, e.g., Imbens and Rubin 2015). Here, we focus on the effect of ad exposure, so our definition of exposure averages over differences in exposure intensity.

Now, we return to the problem of computing ATET. With random assignment,

$$\text{ATET} = E[Y(1)|Z = T, D(T) = 1] - E[Y(0)|Z = C, D(T) = 1].$$

The first quantity is observed: the average outcomes among exposed users assigned to the treatment group. The second quantity is more challenging because the subgroup of consumers with $Z = C, D(T) = 1$ is unobservable. These are control group members who would have been exposed—the *counterfactual exposed* —had they been assigned to the treatment group. We observe the control group $(Z = C, D(C) = 0)$ but do not separately observe the counterfactual exposed consumers $(D(T) = 1)$ or the counterfactual unexposed users $(D(T) = 0)$ within the control group. Next, we propose two solutions for identifying the average counterfactual exposed outcomes: a placebo experiment and ITT.

The first solution is to design an experiment with a placebo treatment that serves to identify the counterfactual exposed users but has no effect of its own. Both the ghost ad and PSA methodologies use this design. Formally, i is assigned to either the treatment group or the placebo control group $Z_i \in \{T, C^P\}$. Now, i's active treatment is the potential outcome $[D_i(T), D_i(C^P)]$. Consumers in the placebo control group receive either the placebo or the null treatment, as in the control group, which we now define as no exposure to either the focal ad or the placebo: $D_i(C^P) \in \{0, P\}$. As before, $D_i(T) \in \{0, 1\}$. In addition, i's potential outcomes are $[Y_i(1), Y_i(0), Y_i(P)]$. Lemma 1 requires two more assumptions to directly identify ATET: (1) the "perfect blind" assumption (Efron and Feldman 1991) means that the focal ad and placebo treatments expose the same consumers; and (2) the "no placebo effect" assumption means that outcomes in the control and placebo groups are the same.

Lemma 1. Given randomization, the exclusion restriction and the stable unit treatment value assumption, as well as for all i:

Assumption A1: (perfect blind) $D_i(C^P) = P$ and $D_i(T) = 1$ or $D_i(C^P) = 0 = D_i(T)$ and

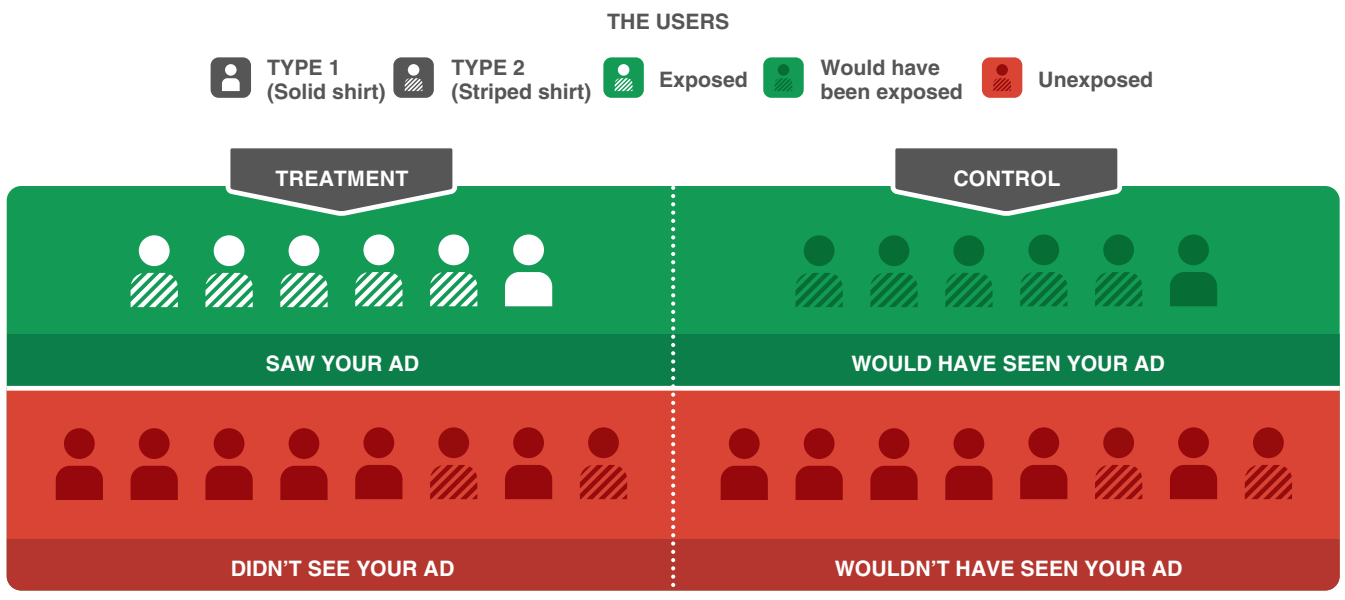
Assumption A2: (no placebo effect) $Y_i(P) = Y_i(0)$,

$$\text{then ATET} = E[Y|Z = T, D = 1] - E[Y|Z = C^P, D = P].$$

Lemma 1 tells us that we can now measure ATET directly with a placebo design (see Web Appendix B for the proof).

Figure 1 illustrates the design of a placebo ad experiment satisfying Assumption A1. Figure 1 categorizes consumers by whether they belong to the treatment or placebo (henceforth control) group and whether the consumers are exposed. We represent consumer heterogeneity by two types of consumers: those who wear striped shirts and those who wear solid shirts. Here, the advertiser is targeting the striped-shirt consumers, who form a majority among the exposed. The distribution of types in Figure 1 across exposed and unexposed is identical

Figure 1
IDEAL EXPERIMENTAL DESIGN



across treatment groups, which illustrates Assumption A1. We evaluate ATET by differencing the outcomes of the exposed and counterfactual exposed—the top half of Figure 1.

At this point, we have kept our ad setting general. The methodologies we discuss apply to settings in which advertising can be varied at the individual consumer level. As such, these methodologies apply to digital media like search and online display advertising, which includes audio and video ads. These methodologies could be more broadly applied by defining the unit of analysis as a geographic region or a household (e.g., addressable television advertising). Next, we discuss two dominant ad experiment approaches: PSAs and ITT.

PSA Approach

One dominant experimental approach is to use PSAs in a control group to identify the counterfactual exposed consumers. To satisfy Assumption A1, the ad platform must deliver the PSAs to the control group in the same way it delivers the focal ads to the treatment group. Then, we can compute the ATET by comparing outcomes among the exposed between treatment and control groups. Letting the PSA play the role of P in Lemma 1, the ad effect is given by

$$(1) \quad \text{ATET}_{\text{PSA}} = \text{E}[\text{Y}|\text{Z}=\text{T}, \text{D}=1] - \text{E}\left[\text{Y}|\text{Z}=\text{C}^{\text{PSA}}, \text{D}=\text{PSA}\right].$$

The ad platform records in a database the identities of consumers and of the ads the platform serves. The PSAs identify the counterfactual exposed consumers, because an analyst can use the ad database to determine which consumers in the control group were served the PSA.

In the experimental context, although PSAs are often charity ads (see, e.g., Hoban and Bucklin 2015; Yildiz and Narayanan 2013), they are more generally neutral ads with an orthogonal call to action to the focal ad to approximate Assumption A2 of "no placebo effect." For our purposes, the PSA approach is equivalent to showing control consumers an ad from an unrelated company (see, e.g., Lewis 2014; Lewis and Nguyen 2015), a blank ad (see, e.g., Bart, Stephen, and Sarvary 2014; Goldfarb and Tucker 2011), or a "house ad" advertising the publisher (see, e.g., Johnson, Lewis, and Reiley 2017; Sahni 2015). As discussed in detail in the "Ghost Ad Approach" section, the PSA approach has drawbacks: PSAs are costly and can violate both Assumptions A1 and A2.

ITT Approach

The second dominant experimental approach is ITT. In 2015, Twitter launched an ITT-based experimentation platform for its large advertisers (Shrivastava 2015). In 2016, Facebook launched an ITT-based experimentation platform which has been used by over 1,000 advertisers (Facebook 2016).

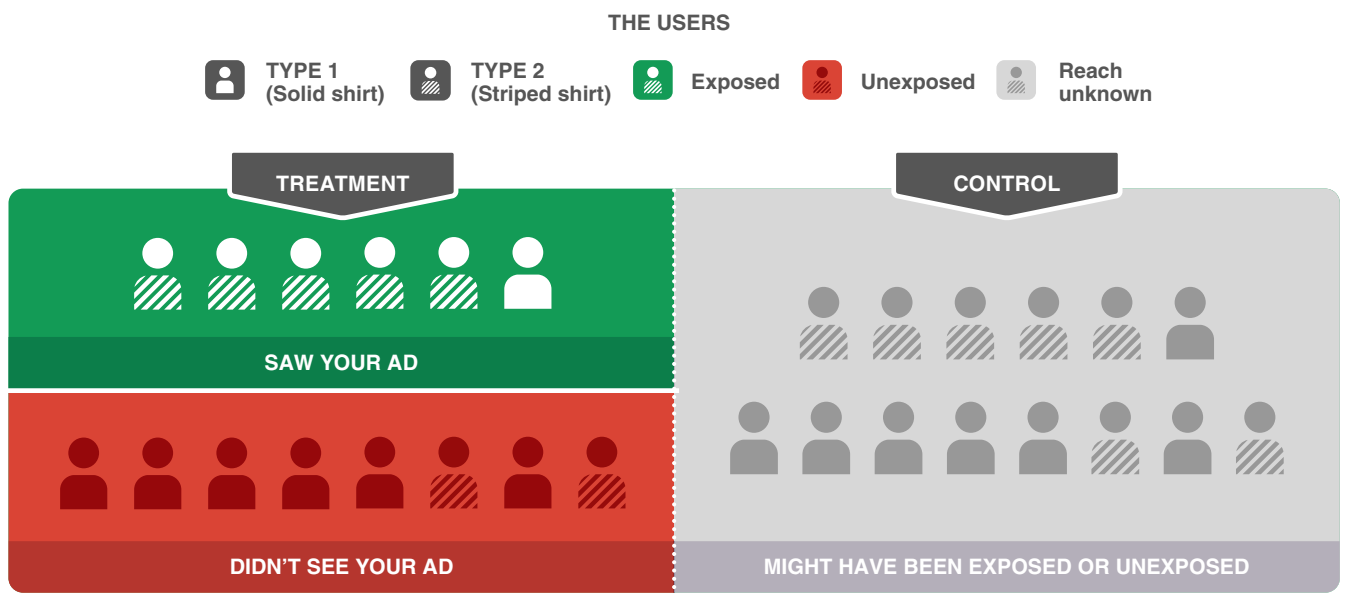
The ITT approach measures the experimental difference among eligible consumers—denoted by the indicator ξ—regardless of ad exposure, given by

$$(2) \quad \text{ITT} = \text{E}[\text{Y}|\text{Z}=\text{T}, \xi = 1] - \text{E}[\text{Y}|\text{Z}=\text{C}, \xi = 1].$$

The ITT approach does not require a placebo treatment like PSAs because ITT ignores whether consumers are exposed. Figure 2 illustrates the logic of using ITT as the counterfactual exposed consumers from Figure 1 are no longer distinguishable. The ITT experimental difference corresponds to the difference in average outcomes between the left (treatment group) and right (control group) halves of Figure 2. Note that the regular, unrefined eligibility $\xi = 1$ is redundant if it includes all consumers who are assigned to treatment or control. More generally, a refined eligibility criterion ξ' is both a precondition to exposure (e.g., consumer is online) and orthogonal to treatment assignment.

The ITT approach serves as a fallback methodology when the PSA, ghost ad, and related methodologies violate Assumption A1. Imbens and Angrist (1994) show that ITT and ATET are related as follows:

Figure 2
INTENT-TO-TREAT EXPERIMENTAL DESIGN



$$\text{(3)} \qquad \text{ATET}_{\text{ITT}} = \frac{\text{ITT}}{\Pr(D=1|Z=T, \xi = 1)}.$$

Intuitively, ITT and ATET are related because ITT's difference must arise from the effect of the ads in expectation, which is the difference among exposed consumers ($D_i(T) = 1$), or ATET. However, the difference among unexposed consumers ($D_i(T) = 0$) is zero, in expectation. As we discuss in the next section, the drawback of this approach is that the indirect ATET_{ITT} estimator is less precise than the direct ATET estimator (e.g., Equation 1).

GHOST AD METHODOLOGY

The purpose of the ghost ad methodology is to generate an indicator variable (GA = P in Lemma 1) that identifies the counterfactual exposed consumers as in Assumption A1. Whereas PSAs identify counterfactual exposed consumers in the ad platform's ad database, the ghost ad methodology instead creates a second database that records the counterfactual exposed consumers.

A ghost ad impression is a log entry in the ghost ad database that records when the focal advertiser's ad *would* have been served to a control group consumer. The ghost ad database thus contains consumer identifiers and an indicator variable for would-be focal ad impressions among all the consumer's served ad impressions. The ghost ad database can be used much like PSAs in the ad database to identify the counterfactual exposed consumers, thereby satisfying Assumption A1. In the ghost ad methodology, PSAs are no longer needed to take the place of the focal ads. Instead, the ad platform allocates ads to control group consumers normally: drawing from the set of available advertisers excluding the focal advertiser. The consumer sees the "next best" ad they would see in the focal advertiser's absence. In practice, consumers see a variety of ads, which may include ads from the focal advertiser's competitors. The ghost ad control group is equivalent to the ITT control group, so Assumption A2 is satisfied. The conditions of Lemma 1 are satisfied, so the ghost ad methodology yields the ATET estimator

$$\text{(4)} \qquad \text{ATET}_{\text{GA}} = \text{E}[Y|Z=T, D=1] - \text{E}\left[Y|Z=C^{\text{GA}}, D=\text{GA}\right].$$

Figure 3 illustrates how the ghost ad methodology works. The upscale shoe manufacturer Christian Louboutin serves as our focal advertiser. Figure 3 shows six ads delivered to an identical consumer in four settings: (1) the treatment group, in which the consumer sees three Louboutin ads and three ads from other advertisers; (2) the control group without control ads (ITT), in which the consumer sees three ads from other advertisers instead of the three Louboutin ads; (3) the control group with PSA ads, in which the consumer sees three PSAs (sea turtle rescue charity ads) in place of the three Louboutin ads; and (4) the control group with ghost ads, in which the consumer sees the same ads as in the ITT setting without control ads, but the ad platform records which three ads are Louboutin's ghost ads in the ghost ad database—depicted by overlaid ghosts.

To determine when the focal advertiser's ad would have been served to a control group consumer, we simulate the ad platform's allocation mechanism. Although these mechanisms can vary, ad platforms usually run an auction, so we henceforth refer to the mechanism as an auction. Figure 4 illustrates how an ad platform processes an ad opportunity in the ghost ad methodology. If the consumer is in the treatment group, the ad auction proceeds normally and selects the winner: either the focal ad or another ad. If the consumer is in the control group, the ad auction selects an ad from the set of participating advertisers excluding the focal advertiser. To identify the counterfactual ad impressions, we also run the ad auction a second time for control group consumers, this time including the focal advertiser among participating advertisers. We only simulate the ad auction in this instance; the selected ad in the simulation has no bearing on the actual ad selected by the auction. We record a ghost ad impression whenever the simulated auction selects the focal ad. Thus, the ghost ad methodology identifies the counterfactual exposed—satisfying Assumption A1—by construction.

We named our methodology "ghost ads" because the two metaphors of "ghost" and "ad" convey the meaning of a ghost ad impression. First, ghosts are invisible. As Figure 3 depicts, ghost ad impressions are invisible to consumers: consumers see the mix of ads they would see absent the focal advertiser. Second, ghosts can "possess" things. As in Figure 3, we imagine that the ghost ads are possessing the ads the consumer sees, but the ad platform knows the possessed ads because they are recorded in the combined ad and ghost ad databases. Third, to the analyst, ads essentially exist as logs in the ad database. In the same way, the ghost ad database logs delineate which consumers are "exposed" to the ghost ads.

The ghost ad methodology should be applied at the ad platform level. To administer the simulated auction, the ad platform must be implementing the focal advertiser's campaign on its behalf. Otherwise, the ad platform will lack the information required to simulate the auction for the control group consumers. In the "Ghost Bid Methodology" subsection, we describe a generalization of ghost ads called "ghost bids" that can be applied when the advertiser's campaign is implemented by another intermediary rather than the ad platform. To use ghost ads, the focal advertiser has no requirements beyond normal business practice—that is, to share consumer marketing outcome information (e.g., website visits) with the platform. An ad platform may want to implement the ghost ad methodology to demonstrate the platform's incremental benefit to advertisers. Ad platforms invest in such services for advertisers because platforms compete with each other and other media for advertisers' ad budgets.

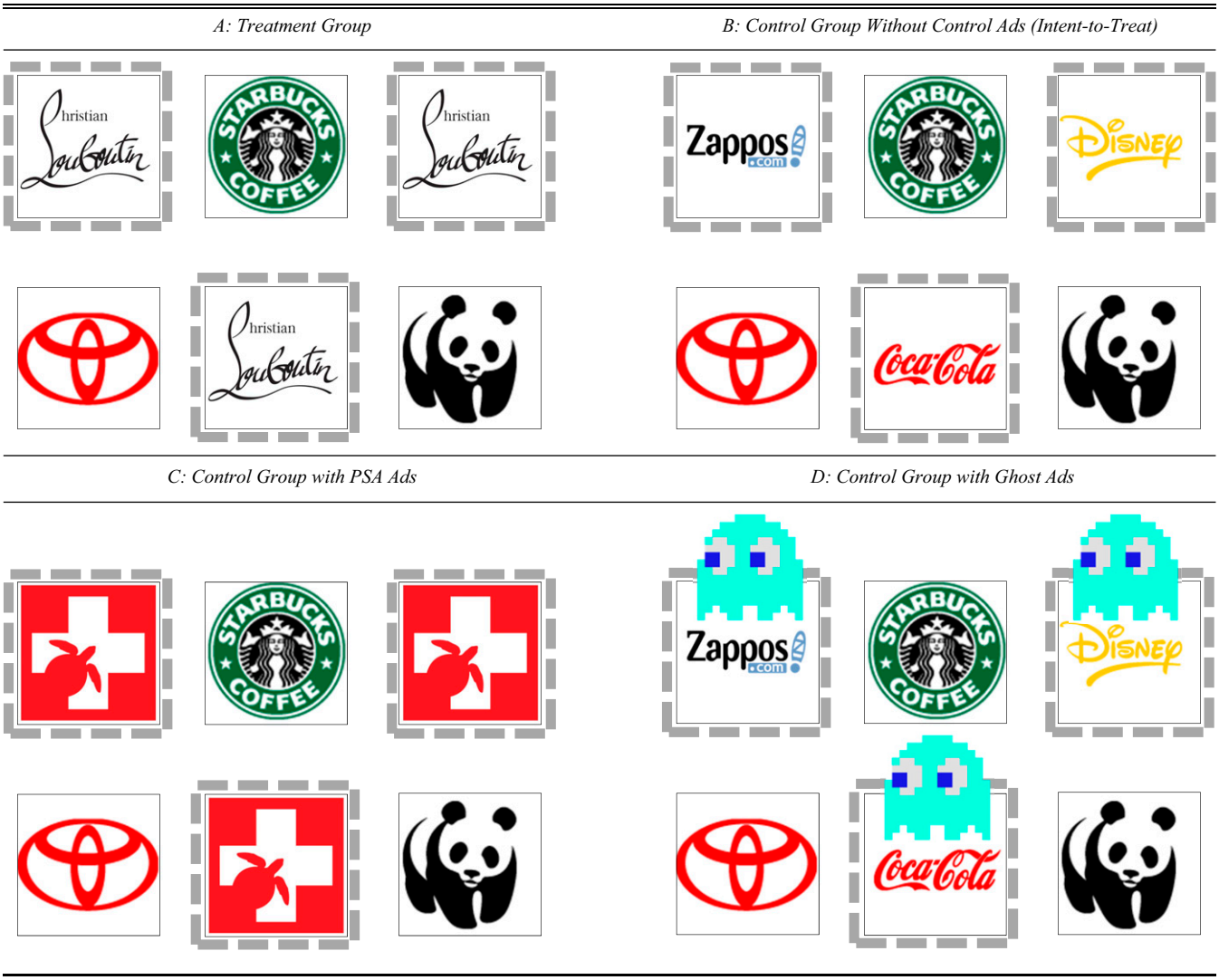
Benefits

The Ghost Ad methodology provides four main benefits vis-à-vis existing experimental methodologies: experimental validity, low cost, the correct strategic baseline, and measurement precision. The following subsections explain these benefits in detail.

Experimental validity. The first drawback of PSAs is that today's advanced ad platforms deliver the PSA and focal ad campaigns differently, such that PSAs can violate Assumption A1. This occurs because modern ad platforms optimize ad delivery by matching each ad to different consumer types. In short, the consumers the platform chooses to expose to the PSAs are no longer a valid control group for the consumers who are exposed to the focal ad (see also Barajas et al. 2016).

Marketers seek to optimize the delivery of their ad campaigns in response to the campaign's performance. In digital advertising, marketers enlist computer algorithms to optimize a campaign outcome like clicks or conversions because low computation cost enables the algorithms to continuously adjust the ad serving during the campaign. These

Figure 3
ADS DELIVERED TO A CONSUMER BY EXPERIMENTAL CONDITIONS



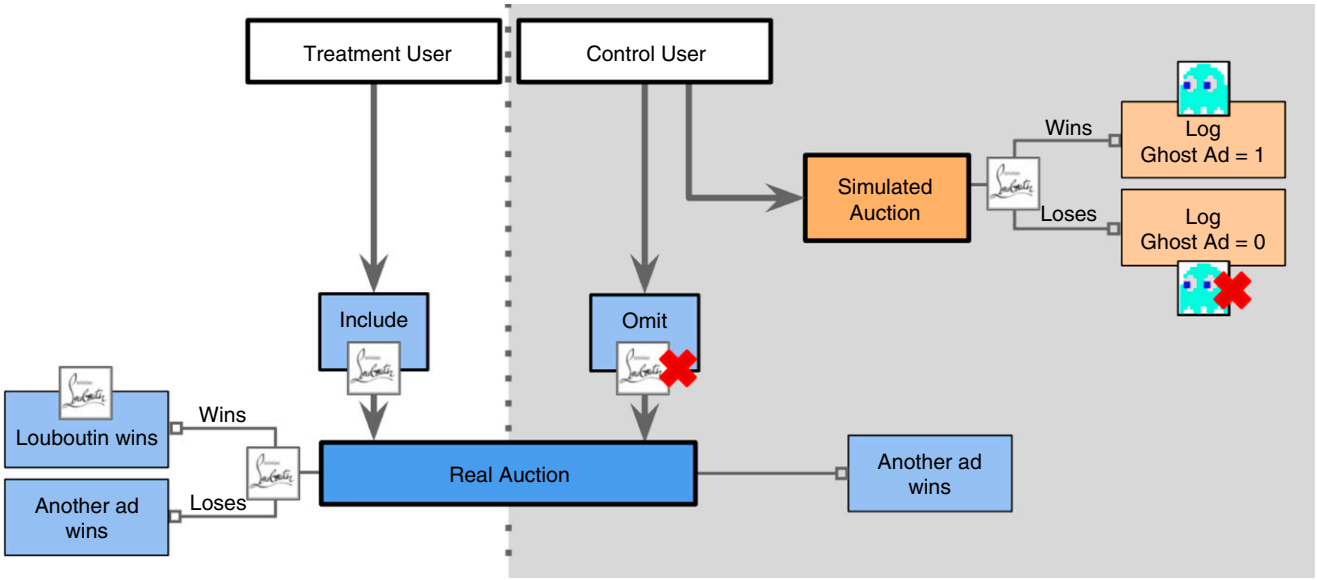
computer algorithms employ machine learning models to predict the probability of clicking—for instance, for different channels (e.g., website), consumer characteristics, and ad creatives. The algorithm then tilts the campaign's delivery toward those channels, consumer types, and creatives the algorithm predicts will have higher click probability.

Automated in-campaign optimization means that the PSAs will violate Assumption A1 and invalidate the ATET estimate (Equation 1) whenever the PSA and focal campaigns are optimized separately. Consider our example Louboutin campaign with sea turtle rescue PSAs. Consumers who are interested in Louboutin shoes will differ from those interested in rescuing sea turtles. Thus, the ad delivery would be unbalanced between the Louboutin ad and PSA. For instance, women who visit fashion sites will see more shoe ads, and men who visit nature sites will see more sea turtle rescue ads. However, baseline Louboutin purchases are higher among fashion-loving women, so comparing conversions between the Louboutin-exposed consumers and the turtle-rescue-exposed consumers will be biased. In fact, this creative-level optimization means that the ad delivery of *any* two distinct creatives, even with the same targeting configuration, will differ. Gordon et al. (2017) document a case in which the $ATET_{PSA}$ estimate yields a negative and significant ad effect whereas the ITT benchmark ad effect estimate is both positive and significant. Figure 5 illustrates that an analysis using PSA campaigns can be biased because the distribution of exposed consumer types will differ between the focal ad and PSA campaigns; in contrast, the ghost ad methodology will balance these types, as Figure 1 shows.

The PSA methodology is therefore confined to primitive ad platforms and nonoptimized campaigns (see, e.g., Johnson, Lewis, and Reiley 2017; Sahni 2015). However, all cost-per-click (CPC) and cost-per-action campaigns are excluded. Considering that two-thirds of online display ad spend uses in-campaign optimization (IAB 2014), the potential application for the ghost ad methodology is significant and growing as ad platforms become more sophisticated.

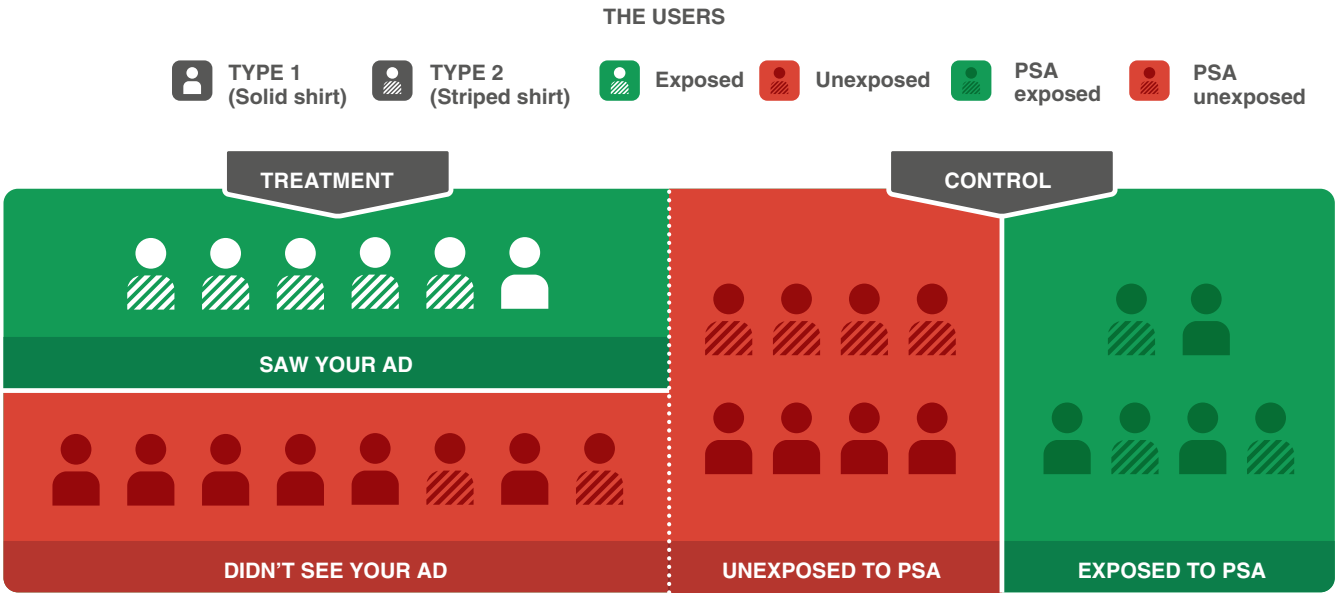
In Web Appendix C, we discuss the nuances of how the PSA and ghost ad methodologies perform as ad platforms become more sophisticated. The most sophisticated ad platforms

Figure 4
GHOST AD FLOW DIAGRAM



can customize consumer-level ad serving during a campaign on the basis of each consumer's actions. Consumer-level ad serving is best exemplified by today's online retargeting campaigns: if the consumer interacts with the advertiser's website or advertisement, the ad platform will infer the consumer's interest and alter ad serving. Often, the ad platform will show interested consumers more of these ads, but the platform may reduce the number of ads, say, after a purchase. The consumer-level campaign feedback creates an endogeneity problem: the consumer's behavior affects his or her exposure intensity. In other words, the consumer-level ad outcomes will mechanically diverge between treatment and control as the campaign causes exposed consumers to take actions; thus, the data that inform the delivery of the ghost ad impressions diverge from the data that inform the delivery of the focal advertiser's impressions. Nevertheless, the control group

Figure 5
BIASED PSA EXPERIMENTAL DESIGN



consumer's first ghost ad impression will still mirror the delivery of the treatment group consumer's first focal ad impression during the experiment because the consumer-level information is equivalent between treatment groups prior to the first impression and the subsequent consumer-level feedback. Ghost ads therefore will satisfy Assumption A1 because they identify the counterfactual exposed, defined as one or more experimental ad exposures; however, ghost ads will not replicate the counterfactual exposure intensity with consumer-level feedback.

Low cost. The second drawback of PSAs is their cost. The ad platform and the focal advertiser must pay for the PSA ad inventory or forgo the revenue from other advertisers that the PSAs displace and compensate publishers. The platform and advertiser must negotiate splitting these costs and coordinate to obtain a suitable PSA. The advertiser must also cover the labor cost of setting up, monitoring, and analyzing the test (Gluck 2011). The ad platform must be configured to handle the PSAs exactly as it handles the focal ad. Such configuration parameters include time period, consumer type, and web page targeting attributes as well as ensuring the ad quantity and budget are proportional to the treatment–control split. Thus, by construction, PSA impressions must cost the same as the focal ad impressions on average, whereas ghost ads are free. Further, small configuration errors are expensive because they can invalidate the experiment by introducing selection bias or complicate any analysis that salvages the experiment. Finally, the costs of obtaining and configuring PSAs persist for additional experiments, making PSAs an obstacle to large-scale, automated experimentation. Thus, although ghost ads greatly reduce costs, we acknowledge that they do not eliminate the opportunity cost of experimentation: the advertiser forgoes the return on investment (positive or negative) from reaching the control group.

Correct strategic baseline. Ghost ads deliver the correct strategic baseline behavior because control group consumers see the ads that they *would* have seen without the experiment. In place of a neutral PSA campaign, the control group sees many advertisers' ads. In our example, we know that the sea turtle rescue association does not run PSAs to rescue Louboutin's target consumers when Louboutin pulls its advertising. Instead, the collection of advertisers that take Louboutin's place could include competitors. From this perspective, the PSAs deliver biased estimates that ignore the strategic effect of the displaced ads, thereby failing the "no placebo effect" Assumption A2 of Lemma 1. The literature documents both positive (e.g., Lewis and Nguyen 2015; Sahni 2016) and negative (e.g., Tuchman 2016) spillovers from competitor ads.

The bias of the PSA estimates (Equation 1) depends on the sign of the net externality of the displaced advertisers. If competing advertisers pull customers away on net, PSA estimates will be too low; they will only capture the "offensive" effect of ads in pulling in consumers and will ignore the "defensive" effect of blocking their competitors from pulling away customers. If the displaced ads instead have positive spillovers, then PSA estimates will be too high because they eliminate this favorable externality. The magnitude of the bias depends on the setting. Strategic considerations are often secondary in display advertising because so many advertisers compete that the overlap in target audiences between direct competitors can be small. Competition is more salient in search advertising, where a few firms compete on a keyword. Note that the experiment can only measure the effect of the ads given the observed behavior of competitors; as with any method, more data must be collected to estimate the net effect if competitors change their strategy.

Consumers in the ghost ad control group see the same mix of ads as they would in an ITT control group. Barajas et al. (2016) and Gordon et al. (2017) concur that this baseline includes the effect of competing ads. However, one unavoidable consequence of both methods is that the presence of the focal campaign in the treatment group will push other competing advertisers toward the control group, where the prices are lower. The focal campaign can mitigate these equilibrium spillovers by reducing the fraction of users assigned to treatment.

We expect this interference problem to be most severe when an advertiser runs another campaign that competes with its ghost ad experiment for the same consumers. As before, the concurrent campaign exposure would be greater in the control than the treatment group, which biases the ATET_{GA} estimator but can be addressed using two-sided noncompliance LATE estimators (Imbens and Rubin 2015, Chapter 24). However, the interference problem can even occur without the advertiser or platform's knowledge if the advertiser purchases ads through multiple intermediaries that do not share information but compete for the same inventory. To avoid cross-campaign interference, advertisers must run one campaign at a time or target each campaign to distinct groups of consumers. For instance, the ghost ad approach can compare the effectiveness of two campaigns (a copy test) by creating three randomly assigned groups of consumers: campaign A treatment, campaign B treatment, and a single control reused twice as a control group with ghost ads for each campaign. The copy test can then deliver the correct strategic baseline for each campaign even with in-campaign optimization.

Measurement precision. The fourth advantage of the ghost ad methodology is measurement precision. Ghost ads share the previous three benefits with ITT. The shared advantage of the ghost ad and PSA placebo approaches is that they enable direct ATET estimation as in Lemma 1. Lemma 2a shows that the direct ATET estimator is more precise than the indirect ATET estimator using ITT estimates from Equation 3.

Lemma 2. (a) For eligibility criterion ξ and exposure rate $\alpha = \Pr[D = 1|Z = T, \xi = 1]$, $\text{Var}(\widehat{\text{ATET}}_{\text{ITT}}) > \text{Var}(\widehat{\text{ATET}}_{\text{GA}})$ and $\text{Var}(\widehat{\text{ATET}}_{\text{ITT}})/\text{Var}(\widehat{\text{ATET}}_{\text{GA}})$ is $O(1/\alpha)$.

(b) Given the more refined eligibility criterion $\xi_2 = 1 \Rightarrow \xi_1 = 1$, $\text{Var}(\widehat{\text{ATET}}_{\text{ITT}_1}) > \text{Var}(\widehat{\text{ATET}}_{\text{ITT2}})$.

The intuition here is that identifying the counterfactual exposed allows the experimenter to prune away consumers who do not see the ad and contribute only noise to the estimator. In particular, ITT includes the experimental differences between exposed consumers and that between unexposed consumers: the latter contribute no ad lift to the estimator but make the estimator less precise. Lemma 2 shows that the gain in precision from the rescaled ITT estimator to the ATET estimator is on the order of the inverse proportion of exposed consumers (see Web Appendix B for the proof).

Lemma 2b takes a broad view that ITT is a class of estimators with eligibility criteria ξ_i that is a precondition for exposure. For instance, ξ_1 may be the set of users targeted by the advertiser, whereas the more refined criteria ξ_2 is the set of

users who are both targeted and online during the campaign. Lemma 2b tells us that the more refined $ATET_{ITT_2}$ estimator will be more precise. Thus, $ATET_{ITT}$ estimators can be viewed on a precision spectrum from all consumers in the world (α small) to the ideal criteria where $\xi = D(T)$ and $\alpha = 1$. If we can refine the eligibility criterion and decrease the proportion of exposed consumers as in the "Ghost Bid Methodology" section, we can improve the precision of ITT.

The precision of the $ATET_{ITT}$ estimator is poor when the proportion of unexposed consumers is high, as is often the case in practice. Typically, the experimenter does not know which control group impressions could have been experimental impressions, as in Figure 2. If all impressions are eligible, the experimenter can only restrict the set of consumers to those who were online to see an ad during the campaign. In this case, the proportion of exposed consumers can be very small (e.g., 3% or less). The problem improves if the experimenter has a predefined list of eligible consumers in the experiment, as in Johnson, Lewis, and Reiley (2017). However, ad platforms often determine eligibility "on the fly" and, therefore, cannot generate such a list of eligible consumers, leaving the set of all online consumers as the only unbiased option.

Johnson, Lewis, and Reiley (2017) show that also identifying the timing of the first counterfactual impression can further improve precision. In particular, the timing data allow the experimenter to filter out the outcome data arising prior to the first ad exposure. Because the outcome cannot be affected before the initial ad exposure, this component of the outcome variable also only adds noise. The experimenter can then define the marketing outcomes as post-exposure rather than during-campaign outcomes. The post-exposure filtering approach is valid when the first experimental ad is delivered symmetrically across treatment groups. In Johnson, Lewis, and Reiley (2017), the ATET estimate's standard error is 31% more precise than their ITT estimate. Defining post-exposure filtered outcomes accounts for a quarter of this improvement in precision. Increasing precision improves the economics of experimentation: in this example, obtaining such precision gains would have required a 110% increase in sample size and cost.

RELATED METHODOLOGIES

The ghost ad methodology can be applied whenever the ad platform controls the ad selection and ad serving processes. Examples include search engine advertising on Bing and Google and streaming services providers like Hulu and Pandora. When the ad platform and another firm jointly determine ad selection or ad serving, new challenges can arise that require related methodologies. The next subsection describes the predicted ghost ad methodology for cases when the ad platform jointly controls ad serving: we implement this approach in the "Empirical Application" section. The following subsection outlines the ghost bid methodology for cases when the ad platform jointly controls ad selection. In particular, ghost bids are a related methodology that an advertiser can implement without the cooperation of the ad platform.

Predicted Ghost Ad Methodology

We first consider the case in which a downstream firm can reject the ad recommended by the ad platform. In online display advertising, the ad platform recommends an ad, which the downstream publisher can reject for many reasons. For example, the publisher may have a secret reserve price that the advertiser's bid does not beat or have hidden exclusions that block certain advertisers. In addition, the ad may be technologically incompatible with the publisher or the user's browser, say, because their system rejects Flash ads. In these cases, the publisher or ad platform will backfill the impression with another ad.

The "downstream publisher refusal" case is incompatible with ghost ads. In the control group, the ghost ad methodology indicates would-be exposed consumers using the simulated auction. In the treatment group, the ad database records whenever the focal ad is *served* rather than whenever the focal ad is *recommended*. Of the recommended and exposed treatment consumers, some proportion x of consumers will not be served a focal ad because the publisher rejected the ad. Thus, the ghost ads methodology yields a biased comparison between the ghost ad–exposed consumers in the control group and the proportion (1 – x) of the recommended and exposed consumers in the treatment group. Moreover, ghost ads cannot work in this context because the ad platform never shares the winner of the simulated auction with the downstream publisher, so the ad platform cannot know when the publisher would reject the focal ad in the control group. To offset these shortcomings, the predicted ghost ad methodology works by simulating the auction for both the treatment group and control group consumers to preserve the symmetric comparison between them.

A predicted ghost ad impression is a log entry in a database that records when the ad platform selects the focal ad in a simulated auction for both treatment and control group consumers. Figure 6 illustrates how the predicted ghost ad methodology handles control and treatment consumers. Both treatment and control consumers enter a simulated auction that includes the focal advertiser—in this case, Louboutin. If the focal advertiser wins the simulated auction, then the ad platform logs this event in the predicted ghost ad database. Unlike the ghost ad database, the predicted ghost ad database logs data for both treatment and control consumers. The real auction then selects ads for the treatment and control consumers to be sent to the publisher, excluding the focal ad from the control group's auction. For both treatment and control consumers, predicted ghost ads $\widehat{D}$ thus approximates the set of exposed and counterfactual exposed consumers denoted by the potential outcome $D(T)$.

Lemma 3 relates the ATET estimator to a LATE estimator. We make the additional Assumption A3 here that $\widehat{D} \equiv \widehat{D}(T)$ is a predetermined binary prediction of $D(T)$ that does not depend on the experimental treatment assignment Z (see Web Appendix B for the proof):

Lemma A3: If $\widehat{D}$ and Z are statistically independent, then

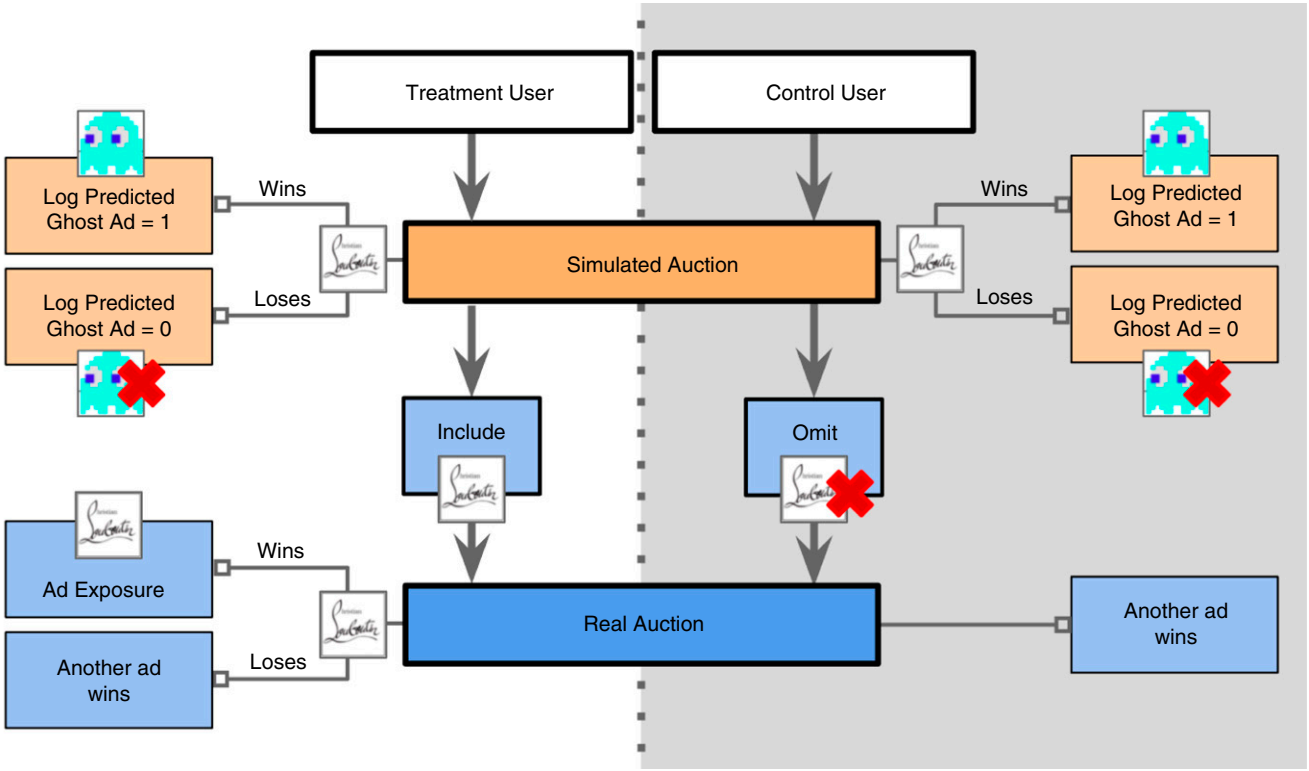
$$ATET = LATE_{PGA} \cdot \Pr\left[\widehat{D} = 1 \middle| Z = T, D = 1\right] + \varepsilon,$$

where

$$\begin{aligned} LATE_{PGA} &\equiv E\left[Y \middle| Z = T, D = 1, \widehat{D} = 1\right] \\ &\quad - E\left[Y(0) \middle| Z = C, D(T) = 1, \widehat{D} = 1\right], \\ \varepsilon &\equiv LATE_{\widehat{D}=0} \cdot \Pr\left[\widehat{D} = 0 \middle| Z = T, D = 1\right], \\ LATE_{\widehat{D}=0} &\equiv E\left[Y \middle| Z = T, D = 1, \widehat{D} = 0\right], \text{ and} \\ &\quad - E\left[Y(0) \middle| Z = C, D(T) = 1, \widehat{D} = 0\right]. \end{aligned}$$

Lemma 3 states that ATET can be written as the weighted sum of two LATEs: the $LATE_{PGA}$ for predicted exposed

Figure 6
PREDICTED GHOST AD FLOW DIAGRAM



consumers and the $\text{LATE}_{\hat{D}=0}$ for "underpredicted" consumers, those not predicted to be exposed. The predicted ghost ad methodology can underpredict exposure; that is, the focal ad can win the real auction but lose the simulated auction. For instance, this can occur if the inputs to the simulated and real allocation mechanisms differ due to engineering constraints like running the two mechanisms sequentially. Underprediction could be serious if the allocation mechanism is probabilistic (see, e.g., Ghosh et al. 2009): when the mechanism allocates a focal impression with probability ρ, a focal ad impression will be underpredicted with probability $1 - \rho$.[2] With the predicted ghost ad methodology, we want the underprediction probability $\Pr[\widehat{D} = 0 | Z = T, D = 1]$ to be zero or low so that the ATET lift is well approximated by the LATE_{PGA} lift. In our empirical application, underprediction is small: only 3.2% of users are exposed but not predicted to be exposed. These underpredicted users see an average of 1.5 ads, whereas predicted exposed users see an average of 20.2 ads, so we expect that $\text{LATE}_{\text{PGA}} \gg \text{LATE}_{\hat{D}=0}$. Because we predict 96.8% of exposed users and predicted-exposed users see 99.8% of the campaign's ads, we expect that the total campaign lift is well approximated by the total lift among predicted-exposed users.

Next, we want to relate LATE_{PGA} to observables. As written previously, LATE_{PGA} requires that we can identify the counterfactual exposed in the control group to compute $E[Y(0)|Z = C, D(T) = 1, \widehat{D} = 1]$, but these consumers are unobserved. Instead, we use an indirect estimator of LATE_{PGA}, which we derive in Lemma 4:

Lemma 4. Given the independence of $\widehat{D}$ and Z,

$$\text{LATE}_{\text{PGA}} = \frac{E[Y|Z = T, \widehat{D} = 1] - E[Y|Z = C, \widehat{D} = 1]}{\Pr[D = 1|Z = T, \widehat{D} = 1]}. \quad (5)$$

The intuition here is that the lift—conditional on predicted exposure—can only arise from exposed consumers (see Web Appendix B for the proof). Equation 5 rescales the experimental difference among predicted exposed consumers by the probability a consumer is exposed conditional on predicted exposure. In our empirical application, this probability is 99.9%, which means the simulator rarely overpredicted treatment in this case. The LATE_{PGA} estimator resembles the relationship between ATET and the rescaled ITT estimator in Equation 3. In fact, LATE_{PGA} is an ITT estimator if predicted exposure is a precondition for exposure, meaning no underprediction occurs.

[2]The underprediction problem can cause situations in which consumers are exposed to a focal ad before their first predicted ghost ad. This, in turn, can create feedback through the platform, which causes the endogenous take-up and ad serving problems discussed in the "Experimental Validity" subsection. Although a focal ad exposure may distort subsequent predicted ghost ad impressions, the underprediction problem is so small here that we do not expect or detect an impact on outcomes prior to predicted treatment.

From the standpoint of measuring the total campaign lift, the predicted ghost ad methodology trades off bias to gain precision. Though $LATE_{PGA}$ is not biased itself, the total campaign lift estimate using only the $LATE_{PGA}$ creates an attenuation bias from ignoring the ad effects on the underpredicted consumers.[3] In contrast, the precision gain with respect to the $LATE_{ITT}$ estimates in Equation 3 can be large depending on the size of the unexposed proportion of consumers (Lemma 2). In our setting, the bias is plausibly small as our $LATE_{PGA}$ estimates capture 99.8% of impressions while the gain in precision is approximately a tenfold reduction in variance. This is an economical trade-off for measuring ad effectiveness.

Ghost Bid Methodology

Now, we consider the case in which the ad platform no longer implements the focal advertiser's campaign. Instead, the ad selection process is jointly determined by the ad platform and another ad intermediary. For instance, online display advertisers often purchase ads through intermediaries such as a demand-side platform (DSP). A DSP has bidding technology that enables it to participate in real-time auctions held by an ad platform called an ad exchange. When the DSP and the ad exchange are separate firms, they may be unable to coordinate on implementing (predicted) ghost ads. For instance, the DSP will not bid on control group users, and therefore, the ad exchange will be unable to simulate the auction with the focal advertiser's bid.

With the ghost bid methodology, the ad intermediary (e.g., the DSP) constructs a database in which it records whenever it submits a bid on the focal advertiser's behalf in the treatment group and whenever the intermediary would do so in the control group. Ghost bids can therefore be used to construct a refined eligibility criterion ξ_{GB} that is independent of Z and enables the advertiser to compute an ITT estimate of the campaign lift (see Equation 2). In this sense, the ghost bid methodology is similar to "exposure logging" in Bakshy, Eckles, and Bernstein (2014) in that ghost bids log activities that are necessary (but perhaps not sufficient) conditions for experimental exposure. Because the ghost bid estimator is an ITT estimator, the disadvantage is a loss of measurement precision (see Lemma 2a). However, ghost bids would be a refined eligibility criterion as in Lemma 2b and would therefore improve the precision of $LATE_{ITT}$ estimates among all eligible consumers. Ghost bids could be further refined if the ad exchange were to post the clearing price of each auction; in practice, ad exchanges do not share this information. Absent this, ghost bids can still be refined by predicting the probability that a bid wins an impression in the treatment group to approximate predicted ghost ads.

The ghost bid methodology has several applications. First, ghost bids enable advertisers and researchers to access most of the benefits of (predicted) ghost ads if the ad platform is unable or unwilling to implement the latter. Second, the ad platform may want to implement ghost bids instead of (predicted) ghost ads because the former may be easier to engineer. Third, the ad platform may want to implement both (predicted) ghost ads and ghost bids, because ghost bids provide a robust fallback to (predicted) ghost ads and can help diagnose implementation problems with (predicted) ghost ads.

EMPIRICAL APPLICATION

Our empirical application features an online display advertising experiment for an online retailer of apparel and sporting goods. A confidentiality agreement prevents us from naming the advertiser, which we will call "Sportsing Inc." for the sake of exposition. Sportsing carries a wide assortment of products, including shoes, jackets, and camping supplies. The firm ran a retargeting campaign that used in-campaign optimization technology to increase consumer transactions. As mentioned previously, PSAs would deliver invalid ad effectiveness estimates in this case. We demonstrate that our predicted ghost ad implementation predicts exposure symmetrically across treatment groups and therefore satisfies Assumption A3. Additionally, we show novel and strong evidence that a retargeting campaign can lift both website visits and purchases.

Experimental Design

Sportsing ran a retargeting display advertising campaign in winter 2014.[4] The experiment assigned 70% of users to the treatment group using a deterministic randomization algorithm that operates on the user's cookie identifier. The campaign delivered an average of 20.2 (SD = 44.0) ad impressions to predicted exposed treatment users on desktop platforms during the two-week experiment. This exposure intensity is moderate relative to the literature. Over the same time period, subjects in Bleier and Eisenbeiss (2015) and Hoban and Bucklin (2015) were exposed to on average 6 ads, whereas subjects in Lewis and Reiley (2014) were exposed to 40 ads. In contrast to previous research, the campaign ran continuously before and after our two-week experiment, and our predicted ghost ad implementation enabled us to start the experiment at our convenience by splitting users into treatment groups. Sportsing ran the campaign on the Google Display Network of over 2 million websites (Google 2015). In the retargeting campaign, Sportsing targeted users who visited its website within the past 30 days, updated on a rolling basis. Eligible users must have browsed a product page, left a product in the online shopping cart without purchasing, or left a product on their online wish list. The 30-day rolling window means that users (re)enter or exit treatment eligibility when they newly visit Sportsing's website or reach the 30-day limit without visiting. In addition, users who purchase from Sportsing during the campaign are ineligible to see the experimental ads. Sportsing used in-campaign optimization to reduce the cost per conversion, where Sportsing fixed the conversion to be a purchase among users who clicked on the ad during the campaign. While Sportsing also values purchases by users who do not click, this setup is used in industry to attribute a purchase to the last-clicked ad impression. To evaluate the campaign's effectiveness, we obtained data on both purchases and website visits at Sportsing using pixels.

For the campaign, Sportsing employed dynamic retargeting display ads that featured the products that the individual user had viewed on the advertiser's site. Figure 7 shows an

[3]While we could estimate $LATE_{\hat{D}=0}$ analogously to Equation 5, our estimates would be very imprecise because the underpredicted consumers represent a tiny fraction of eligible consumers. The logic is the same as Lemma 2.

[4]Sportsing was the primary research and development partner through the development of predicted ghost ads. This was the first well-powered experiment among the two early clients running tests.

example of a dynamic remarketing creative, which is similar to Lambrecht and Tucker's (2013) dynamic retargeting creative. Sportsing's ad features its own logo as well as two product photos against a neutral background. The flash ad rotates through featured products every few seconds. The user could click on the ad to go to Sportsing's website or to the featured product's page. The user could also click on arrows in the ad next to the product pictures to scroll backward or forward through the product photos. When the user moused over a product photo, that product's brand name would appear below the product photo. The campaign also delivered some smaller, text-based ads headed by a hyperlink that was followed by short lines of text all alongside a small product picture and the product price. In all cases, the ads do not promote a sale.

Experimental Validation

Experiments must demonstrate the validity of the randomization by showing that the treatment and control groups are equivalent prior to exposure. To do so, we test for differences in the subjects' characteristics and pretreatment outcomes. We perform these checks on the set of predicted exposed users, which also validates that our predicted ghost ad implementation satisfies Assumption A3. We further validate that our implementation satisfies A3 by testing for differences in the delivery of the first predicted impression across treatment groups. Table 1 lists p-values, which show that the experiment passes all the tests with respect to user characteristics, first impression delivery, and pre-experimental outcomes.

To begin, we verify that the user characteristics are equivalent in the treatment and control groups conditional on predicted ghost ad exposure. The study includes 566,377 predicted-exposed users, of which 396,793 were in the treatment group and 169,584 were in the control group, implying a 70.06% treatment-group share of exposed users relative to the expected 70% share ($p = .34$, two-sided binomial test). We know little about users who are browsing sites across the Google Display Network, but we can infer the user's location from their IP address. In Table 1, we compare the distribution of two categorical variables that indicate the user's country and city location. Chi-squared tests fail to reject ($p = .85$ and $p = .50$).

Figure 7

RETARGETING CREATIVE RESEMBLING THOSE IN EXPERIMENT

In Table 1, we test the predicted ad exposure variables for users' first predicted impression. As discussed previously, a retargeting campaign will only replicate the delivery of the first predicted ad across the treatment and control groups. Thus, the characteristics of all users' first predicted ghost ads should be equal because the system has equivalent information before each user's first ad impression, regardless of treatment assignment. Here, the first treatment and control predicted ghost ads are delivered symmetrically across treatment groups for the 539 websites in the campaign ($p = .49$). Moreover, both groups see the same distribution of the first predicted ad's creative format: flash or text formats as well as different ad shapes and sizes ($p = .68$). We also find no significant differences for t-statistic tests of equality of means for the first predicted ad's CPC ($p = .38$), predicted conversion probability ($p = .10$), predicted click probability ($p = .47$), and predicted cost per impression ($p = .10$). The ad platform calculates these variables for each impression as they affect the platform's delivery decisions. We were able to present the treatment and control means for these continuous variables only, which we show in Table 1. In summary, we see no significant differences in the characteristics of the first predicted impression.

Finally, we test the predicted ghost ad system's delivery with respect to precampaign outcomes. In Table 1, we see no differences in site visits ($p = .92$), transactions ($p = .76$), and sales ($p = .75$) across groups during the 30 days before the campaign. We also test for pre-exposure differences in outcomes between the start of the campaign and the first predicted exposure; we want to filter out these data to increase the precision of our estimates. Again, Table 1 shows no significant differences ($p = .33$, .21, and .11), instilling confidence that our experimental estimates represent the ad lift and not a failure of our predicted ghost ad implementation.

Although the first predicted ghost ad should be symmetric across the treatment and control groups, subsequent impressions should not. As discussed previously, the subsequent predicted ghost ad exposures are distorted by feedback from exposed users responding to the ads, which in turn alters the ads they receive. For instance, users in the experiment are dropped from the ad campaign once they purchase. This means that the incremental users in the treatment group—whom the ads *cause* to purchase—will receive fewer predicted ads than their (unidentifiable) counterparts in the control group. To examine this, we construct user-level variables that average over the user's predicted impression characteristics (e.g., CPC). As expected, Table 1 reveals significant differences across treatment groups in user-level average predicted impression variables. Indeed, a lack of difference between subsequent treatment and control impressions might be symptomatic of an ineffective campaign—one that fails to produce the mechanical feedback of a user-optimized ad platform.

Table 1
EXPERIMENTAL VALIDATION FOR PREDICTED GHOST AD EXPOSED USERS

	Treatment/Control Test of Equality			
Demographics	Exposed Users			
Country	$\chi^2_{12} = 7.1$		$p = .85$	
City	$\chi^2_{1,095} = 1,094.6$		$p = .50$	
Predicted Ghost Ads	First Impression[a]		Average Impression[c]	
Creative Shown	$\chi^2_{47} = 3.748$	$p = .68$	—	—
Publisher Website	$\chi^2_{538} = 532.2$	$p = .49$	—	—
Predicted Click-Through Rate[b]	t = .73	$p = .47$	t = 10.42	$p = .00$
Treatment Mean/(SD)	.57%	(.79%)	.48%	(.52%)
Control Mean/(SD)	.57%	(.80%)	4.46%	(.51%)
Predicted Conversion Rate[b]	t = 1.63	$p = .10$	t = 8.77	$p = .00$
Treatment Mean/(SD)	2.71%	(2.63%)	2.76%	(2.41%)
Control Mean/(SD)	2.70%	(2.61%)	2.70%	(2.35%)
Predicted Cost per Impression[b]	t = 1.65	$p = .10$	t = 8.41	$p = .00$
Treatment Mean/(SD)	$2.944	($5.668)	$5.527	($3.338)
Control Mean/(SD)	$2.913	($6.629)	$5.414	($5.070)
Cost per Click[b]	t = .43	$p = .38$	t = 3.71	$p = .00$
Treatment Mean/(SD)	$.637	($.714)	$.682	($.630)
Control Mean/(SD)	$.636	($.718)	$.675	($.626)
Pre-Treatment Outcomes	30 days prior to experiment		During experiment prior to exposure	
Site visits	t = .31	$p = .76$	t = 1.27	$p = .21$
Transactions	t = −.11	$p = .92$	t = −.97	$p = .33$
Sales	t = −.31	$p = .75$	t = −1.62	$p = .11$

[a]First predicted ghost ad impression following the start of the experiment. For ongoing campaigns, users may have already seen impressions prior to the start of the experiment.

[b]These quantities are calculated by the Google Display Network's internal ad serving systems.

[c]Not expected to be equivalent between treatment and control. Averages taken across impressions at the user level.

Notes: Tests are between 396,793 predicted exposed users in the treatment group and 169,584 predicted exposed users in the control group. Categorical variables employ $\chi^2_{d,f}$ and continuous variables employ t-tests for difference in means.

Results

Now we apply our predicted ghost ad methodology to measure the effect of Sportsing's retargeting ads on sales and site visits. We use the predicted ghost ad methodology to estimate the $LATE_{PGA}$ from Equation 5 on the predicted exposed users. Table 2 lists both the lift among predicted-exposed users ($LATE_{PGA}$) and eligible users (ITT) for robustness.

Our predicted ghost ad $LATE_{PGA}$ estimates yield highly significant evidence that Sportsing's ads increase both site visits and sales. Following Johnson, Lewis, and Reiley (2017), we only include sales following the first-predicted ad exposure, rather than from the start of the experiment, to increase the precision of our $LATE_{PGA}$ estimates. Table 2 shows that Sportsing's ads increase site visits by 62,756, or 17.2%, and sales by $105,030, or 10.5%. The corresponding t-statistics and p-values are 13.64 ($p < 10^{-15}$) for site visits and 3.51 ($p < 10^{-3}$) for sales. We also examine how the ads affect the extensive margin in terms of the number of site visitors, transactors, and transactions. In Table 2, we see the ads cause 19,066 incremental visitors (26.6%), 1,015 incremental transactors (12.1%), and 1,131 incremental transactions (11.9%). These results are even more significant than their intensive-margin counterparts with t-statistics of 41.81 ($p < 10^{-15}$) for visitors, 6.02 ($p < 10^{-8}$) for transactors, and 5.54 ($p < 10^{-7}$) for transactions.

Although our predicted ghost ad implementation passes all the validation tests above, we also compare our $LATE_{PGA}$ estimates to the ITT estimates for robustness. Given the dynamic nature of the campaign eligibility, we do not know the number of eligible users in the experiment. To measure ITT, we compute the experimental difference among all of Sportsing's site visitors, which the deterministic randomization algorithm can sort into treatment and control users. Because the number of eligible users is unknown, we use a conservative approximation to compute the ITT standard errors (see the Appendix for details). To make the ITT and $LATE_{PGA}$ estimates comparable, we examine the total campaign lift rather than the user-level average lift because the latter is ill-defined for ITT with unknown N. Table 2 confirms that the ITT and $LATE_{PGA}$ estimates are close and that a Hausman test for each outcome does not reject the $LATE_{PGA}$ estimator.[5] Though our conservative ITT standard errors approximation make the Hausman test less strict, the test has high power for visits and visitors but low power for sales.[6] Given the Hausman tests in Table 2 and the validation checks in Table 1, we can be confident that our predicted ghost ad implementation delivers valid ad effectiveness estimates.

[5]The ITT estimate and post-predicted-exposure $LATE_{PGA}$ estimates for site visitors are significantly different because extensive margin outcomes are not additively separable over time. Therefore, we report Hausman tests for the during-campaign $LATE_{PGA}$, which reveal no significant differences from the during-campaign ITT estimates for both site visitors and transactors.

[6]A power calculation reveals that the respective Hausman tests have 90% power to reject the following absolute differences between the ITT and $LATE_{PGA}$ estimates at the 5% two-sided level: 6,149 visitors (3.18% of the PGA control), 46,558 visits (12.78%), 1,374 transactors (15.90%), 2,093 transactions (22.09%), and $460,137 in sales (45.36%).

Table 2
AD EFFECTIVENESS RESULTS FOR DYNAMIC REMARKETING CAMPAIGN

	1	2	3	4	5
	Site Visitors	*Site Visits*	*Transactors*	*Transactions*	*Sales*
Predicted Ghost Ad LATE Estimates: Post-First Predicted Impression Outcomes					
Treatment	90,835	426,967	9,408	10,605	$1,110,048
	(265)	(2,632)	(96)	(117)	(17,959)
Control[a]	71,769	364,211	8,393	9,474	$1,005,018
	(371)	(3,774)	(139)	(168)	(23,977)
Difference	19,066	62,756	1,015	1,131	$105,030
	(456)	(4,601)	(169)	(204)	(29,957)
t-Statistic	41.81	13.64	6.02	5.54	3.51
% Lift	26.6%	17.2%	12.1%	11.9%	1.5%
Average difference[b]	.048	.158	.0026	.0029	$.26
ITT Estimates During Compaign Outcomes					
Difference	12,238	69,821	1,274	1,861	$265,165
	(1,800)	(13,751)	(419)	(617)	(131,540)
t-statistic	6.80	5.08	3.04	3.02	2.02
Hausman test (p)	.61[c]	.59	.47[c]	.21	.21
Relative variance of ITT	15.6	8.9	6.2	9.1	19.3

[a]Control group total outcomes rescaled by 7/3 to match treatment groups in 70%/30% treatment split.

[b]Average difference is lift divided by 396,793 predicted exposed users in the treatment group.

[c]Hausman test compares during-campaign ITT estimates above with during-campaign (rather than post-predicted exposure) $LATE_{PGA}$ estimates (see footnote 5): the $LATE_{PGA}$ estimates become 13,106 (SE = 575) for visitors and 998 (SE = 171) for transactors.

Notes: Results employ experimental difference regression estimator. Robust standard errors are in parentheses; ITT standard errors use the conservative approximation $\text{Var}(\sum y_i) \gg \sum y_i^2$ because the number of users who could purchase is unknown (see the Appendix for details).

Johnson, Lewis, and Nubbemeyer (2017) provide evidence that the predicted ghost ad methodology was implemented successfully by examining 432 such experiments at the Google Display Network. Johnson, Lewis, and Nubbemeyer show that the Hausman test rejects the $LATE_{PGA}$ estimates at the 5% level in only 5% of cases, which is consistent with false positives. The Sportsing lift estimates are close to the median lift estimates in Johnson, Lewis, and Nubbemeyer: 17% lift in site visits and 8% lift in conversions (here, transactions). In Web Appendix D, we present a representative sample of eight studies drawn at random from those in Johnson, Lewis, and Nubbemeyer.

Not only do our results demonstrate for the first time that retargeting can be effective, but they do so with greater statistical significance than most field experiments in the ad effectiveness literature. A survey of marketers found that 85% spend at least 10% of online ad budget on retargeting (AdRoll 2014). Industry studies of retargeting present a rosy view of its effectiveness, claiming 70%–1,600% improvements (Econsultancy 2014; Hunter et al. 2010; PriceWaterhouseCoopers 2011). Our result is important because retargeting is fraught with endogeneity problems that make its true effectiveness controversial: retargeted users are likely to purchase without any ads because they are a self-selected group that has demonstrated interest in purchasing. In fact, retargeted ads could even reduce sales if the ad's overt tracking provokes reactance in users. By comparing retargeted users with a valid holdout group, we can show that this retargeting campaign creates a large increase in both site visits and sales. This baseline comparison with a control group differentiates our results from Lambrecht and Tucker (2013) and Bleier and Eisenbeiss's (2015) retargeting studies. These studies instead compare retargeting campaigns that feature more or less personalized information regarding the focal product.

The (predicted) ghost ad methodology changes the economics of obtaining ad effectiveness information for advertisers. In our study, a midsize online retailer obtains highly significant results despite having a budget of only $30,500. If Sportsing had to pay for PSAs for the 30% control group, the experiment's cost would rise by 43%. Beyond the cost reductions, the ghost ad methodology changes the kind of experiments advertisers would want to run. For instance, Sportsing can obtain the same statistical power with a 30/70 treatment control design rather than the current 70/30 design and save 58% of its ad budget. This means advertisers can run experiments that treat a small proportion of eligible users and then roll out successful campaigns to the whole group. Additionally, Sportsing can learn more with less money by running a concentrated test that increases average spend per user while decreasing the proportion of exposed users. Sportsing can obtain the same statistical power as the original test by running a concentrated test that doubles the ad spend per user while shrinking the treatment group to a 6/94 design, if the ad effect is proportional to ad spend.[7] Now, costs fall 83% to only $5,000—almost an order of magnitude lower than the $45,000 cost of a PSA test. Concentrated tests are like "accelerated failure tests" in that they enable advertisers to discover successful ad campaigns at low cost, which again can be rolled out more broadly.

Thus, the (predicted) ghost ad methodology improves on Lewis and Rao's (2015) pessimistic outlook on measuring the

[7]Denote the average experimental difference by δ and the number of users by N. Assume the variance of the outcome σ^2 is equal across treatment groups. For the t-statistic $= \delta/\text{SE}(\delta)$ to be the same when the treatment effect doubles, we find that the proportion of users in the treatment group p = 5.6% solves $2\delta/\frac{\sigma}{\sqrt{N_p(1-p)}} = \delta/\frac{\sigma}{\sqrt{N \times .3 \times .7}}$. Although we are assuming no ad wear-out (i.e., diminishing returns) from doubling the ad spend, this is consistent with the findings of Lewis (2014) and Johnson, Lewis, and Reiley (2017).

returns to advertising; they use a meta-study to argue that the setting's statistical power problem is so severe that experiments require many millions of user-week observations to draw conclusions. The problem results from the fact that the effects of advertising can be orders of magnitude less than the noise in the data. Here, we obtain strong results because the predicted ghost ad methodology decreases the variance in the estimates relative to ITT. The ratios of these estimates' variances in Table 2 range from 5.9 to 16.4, which indicates that experiments using only ITT estimates would need to be an order of magnitude larger to reach comparable statistical confidence.[8]

Some technological limitations will attenuate our ad effectiveness estimates. For instance, consumers refresh their cookies from time to time. Once a cookie is deleted, the user's subsequent activity will be missing. This attenuates our estimates because we miss the incremental consumers who transact after seeing an ad but change their cookies in the interim. Moreover, a cookie-switching consumer that reenters the experiment could switch treatment groups, further attenuating the measured ad lift. The median cookie age in our data is approximately 3.5 months, so this problem does not affect the majority of users in our sample. Similarly, each computer, browser, tablet, or mobile device has its own unique anonymous cookie that is independently randomized, so a consumer could receive different treatments, again attenuating our estimates (Coey and Bailey 2016). In addition, if a user's ad exposures are linked to one cookie but his or her purchases are linked to another, our estimates will be further attenuated. That said, Sportsing's online retail and advertising strategies were largely desktop-centric during the experiment. Given the attenuation biases induced by these technological limitations, we interpret our ad lift estimates as lower bounds.

IMPLEMENTATION CHALLENGES

The predicted ghost ad methodology presents several implementation challenges. In this section, we discuss some lessons learned from implementing the methodology within the Google Display Network and a possible extension to ad viewability.

From the perspective of the ad platform, the fixed and marginal costs of the ghost ad approaches are low. The development costs can be low as well because these approaches—notably the simulation step—leverage the ad platform's existing capabilities. Though the ghost ad approaches require computational resources to simulate and record ghost ad events, computational costs in online advertising are a small fraction of ad revenues. The more important consideration for the ad platform is opportunity cost. In the short run, advertisers may reduce their campaign budget to only reach the fraction of consumers in treatment. In the long run, ad budgets will rise or fall as a function of the incremental effect of the ads. Regardless, competition between platforms could push them to offer advertisers the ability to experiment.

The fundamental challenge with implementing the proposed methodologies is to ensure symmetry between treatment groups—else the experimental comparison would be biased. Perfect symmetry is elusive because online display ad platforms are complex and ever evolving. This means that our predicted ghost ad implementation and its associated experimental analysis pipeline must be monitored and updated as new versions of the ad platform code are deployed. As mentioned previously, the predicted ghost ad methodology is robust to the choices of a downstream publisher. For the same reason, the methodology is robust to some unexpected consequences of the ad platform's complexity and evolution.

Some challenges arise from the interactions between multiple ghost ad experiments. For instance, suppose that the ad platform is concurrently running ad experiments for Louboutin and IKEA. Now, suppose that consumers overlap between experiments so that some consumers are in both IKEA and Louboutin's control group. If IKEA wins the simulated auction, then an IKEA predicted ghost ad is logged but not one for Louboutin. If Louboutin could have won the auction in the absence of IKEA, then Louboutin's ghost ad impression is censored. Instead, the simulated auction—from Louboutin's vantage point—should exclude IKEA since IKEA's ad can never win in IKEA's control group. Our "auction isolation" solution eliminates cross-campaign externalities by running separate simulated auctions for each overlapping ghost campaign. In each isolated auction, the set of participating advertisers is a single experimental ad campaign and the competing advertisers who could actually win the campaign. Thus, the predicted ghost ad database can assign a consumer's single ad impression to multiple experimental campaigns.

Many implementation challenges arise from the combinatorial nature of the ad-allocation auctions employed by ad platforms. For example, a single display ad slot on Google can be won by either a single display ad or by multiple separate text-based ads. Thus, the predicted ghost ad system must record one or multiple winners. Some ad platforms also enable advertisers to purchase multiple impressions on a page at the same time or to ensure that their ad appears only once on a page. Ad platforms frequently use combinatorial auctions to implement these features (Candela, Bailey, and Dominowska 2014). This allows ads—including ghost ads—to form coalitions that can alter the number of experimental ads on a page and complicate the prediction process. Although auction isolation avoids collisions between ghost ads here, the simulated auctions must be implemented over all impressions on a page to work. These complexities necessitate the simulated auction step: ghost ads cannot simply allocate the ad to the "second highest bidder" whenever the focal advertiser wins the real auction.

The ad platform can also modify the predicted ghost ad methodology to record information on ad viewability. A viewable impression is an ad that appears in the consumer's field of view for a certain length of time and is designed to discount ads that appear below the fold or for a split second. Marketers may want to measure the impact of viewable ads considering that, logically, nonviewable ads have little or no effect. However, the ad platform can only record viewability information for the ads it serves and not those served by another intermediary. Thus, while the ad platform has viewability information for the focal ads, the viewability data are incomplete for ads shown to the control group. The ad platform can solve this problem by identifying the predicted displaced ad, the ad that is predicted to be served to the treatment group in the absence of the focal ad. To operationalize this, the ad platform can run simulated auctions both with and without the focal ad for all consumers and record the predicted ghost ad and predicted displaced ad. In this way, the ad platform can split the predicted displaced ads (conditional on

[8]The experiment also enjoyed better statistical power from a larger (30%) control group, concentrated average ad frequency (20 impressions per user), and directly linkable online-only purchase channel. However, this retailer typifies those in Lewis and Rao (2015) with a similar coefficient of variation on sales of $\sigma/\mu = 10.1$ during this two-week campaign.

predicted ghost ad exposure) by whether they record viewability and do so symmetrically across treatment groups.

CONCLUSION

Our ghost ad and related methodologies promise to improve ad effectiveness measurement by changing the economics of experimentation. With experiments, the advertiser's learning can be quantified by the precision of the ad effectiveness estimate. Our application demonstrates that advertisers can learn just as much from a ghost ad test as from a PSA test while spending an order of magnitude less. These savings arise from (1) eliminating the cost of PSAs, (2) assigning more consumers to the control group without ad costs rather than to the treatment group, and (3) using concentrated tests with high spending on small treatment groups to increase the average ad lift. Our application also shows that learning with ghost ads is much cheaper than ITT tests: a comparable ITT experiment would need an order of magnitude larger budget and sample size to compensate for ITT's imprecision. In both cases, the advertiser learns an order of magnitude more per dollar spent on the Ghost Ad experiment.

The improved economics of measuring ad effectiveness should encourage more advertisers to not only start experimenting and but also make experimentation routine. Already, hundreds of advertisers have used Google's early implementation of the methodology, which delivers hundreds of millions of experimental ad impressions daily. As more advertisers run experiments, we foresee that three kinds of experimental tests will emerge. First, advertisers will use monitoring tests with small control groups to routinely account for both the short- and long-term effects of advertising. Second, advertisers will use concentrated tests with high spending on small treatment groups to optimize a campaign's return on investment. Third, advertisers will use explore-exploit tests, in which advertisers explore the effects of different campaigns on small groups of consumers before exploiting the best campaigns on the rest.

Ghost ads and related methodologies also present opportunities for researchers. By reducing the cost of running experiments, researchers will more easily convince marketers to run field experiments. Widespread experimentation would enable academics and practitioners to refine their ad models. In follow-up research, we examine over 400 predicted ghost ad experiments averaging 4 million users each (Johnson, Lewis, and Nubbemeyer 2017). We show in our meta-analysis that predicted ghost ad experiments can examine important marketing questions like measuring the in-campaign and carry-over effects of advertising on consumer behavior. We hope that other meta-studies of ghost ad experiments will shed new light on ad attribution and ad stock models.

The ghost ad methodology keeps pace with modern ad platforms' performance-optimizing features that now account for two-thirds of online display ad spending (IAB 2014). What is more, ghost ads can be applied to other forms of advertising beyond online display ads. Search ads are natural candidates because they are sold using similar digital infrastructure. Other media could follow, considering that programmatic ad buying and ad measurement are spreading to television advertising. In the future, we predict new advertising media experiments will employ the ghost ad methodology to improve the theory and practice of marketing.

APPENDIX: ITT STANDARD ERROR APPROXIMATION

Our ITT estimates compare the total difference in the marketer's outcome y between the N_T consumers who were eligible to be in the treatment group and the N_C users who were eligible to be in the control group. The total ITT lift is then given by

$$\text{Total ITT} = \sum_{i=1}^{N_T} y_{i,T} - \frac{7}{3}\sum_{i=1}^{N_C} y_{i,C},$$

where we rescale the total control group outcome because the treatment-control group split is 70/30. Now, N_T and N_C are unknown and represent all users who could engage in the marketing outcome y with positive probability. Instead, we observe $N^{\emptyset}$ non-zero outcomes $y^{\emptyset}$, which suffices to measure the sum $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N^{\emptyset}} y_i^{\emptyset}$. Now, we want to compute the standard error for the total ITT lift but need to approximate this because an exact calculation requires that we know N:

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{N} y_i\right) &= N \times \text{Var}(y) = N\left[\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2\right] \\ &= \sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2.\end{aligned}$$

We derive a conservative approximation for this variance $\text{Var}(\sum_{i=1}^{N} y_i) \approx \sum_{i=1}^{N^{\emptyset}} (y_i^{\emptyset})^2$ using the following logic:

$$\text{Var}\left(\sum_{i=1}^{N} y_i\right) \le \sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N^{\emptyset}} \left(y_i^{\emptyset}\right)^2,$$

where the squared sum of non-zero outcomes $\sum_{i=1}^{N^{\emptyset}} (y_i^{\emptyset})^2$ is observed. The quality of our approximation is best if $\frac{1}{N}(\sum_{i=1}^{N} y_i)^2$ is small. This is realistic in our setting because N is much higher than the $N^{\emptyset}$ consumers who take the marketing action. Sparse outcomes mean that $E[y]$ is close to zero, but $\text{Var}(y)$ is large relative to $E[y]$; hence $\sum y^2 \gg \sum y$.

REFERENCES

AdRoll (2014), "State of the Industry: A Close Look at Retargeting and the Programmatic Marketer," technical report, AdRoll.

Bakshy, Eytan, Dean Eckles, and Michael S. Bernstein (2014), "Designing and Deploying Online Field Experiments," in *Proceedings of the 23rd International Conference on World Wide Web*. New York: Association for Computing Machinery, 283–92.

Bakshy, Eytan, Dean Eckles, Rong Yan, and Itamar Rosenn (2012), "Social Influence in Social Advertising: Evidence from Field Experiments," in *Proceedings of the 13th ACM Conference on Electronic Commerce*. New York: Association for Computing Machinery, 146–61.

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), 465–83.

Bart, Yakov, Andrew T. Stephen, and Miklos Sarvary (2014), "Which Products Are Best Suited to Mobile Advertising? A Field Study of Mobile Display Advertising Effects on Consumer Attitudes and Intentions," *Journal of Marketing Research*, 51 (3), 270–85.

Blake, Thomas, Chris Nosko, and Steven Tadelis (2015), "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment," *Econometrica*, 83 (1), 155–74.

Bleier, Alexander, and Maik Eisenbeiss (2015), "Personalized Online Advertising Effectiveness: The Interplay of What, When, and Where," *Marketing Science*, 34 (5), 669–88.

Candela, Joaquin Q., Michael Bailey, and Ewa Dominowska (2014), "Machine Learning and the Facebook Ads Auction," presentation in Joint Session for EC, NBER and Decentralization on CS and Economics, EC14, Palo Alto, CA.

Coey, Dominic, and Michael Bailey (2016), "People and Cookies: Imperfect Treatment Assignment in Online Experiments," in *Proceedings of the 25th International Conference on World Wide Web*. New York: Association for Computing Machinery,

Econsultancy (2014), "Display Retargeting Buyer's Guide," technical report, Econsultancy.

Efron, Bradley, and D. Feldman (1991), "Compliance as an Explanatory Variable in Clinical Trials," *Journal of the American Statistical Association*, 86 (413), 9–17.

Facebook (2016), "How We're Making Ad Measurement More Insightful," Facebook for Business News.

Ghosh, Arpita, Preston McAfee, Kishore Papineni, and Sergi Vassilvitskii (2009), "Bidding for Representative Allocations for Display Advertising," in *Internet and Network Economics*. Berlin: Springer, 208–19.

Gluck, M. (2011), "Best Practices for Conducting Online Ad Effectiveness Research," technical report, Internet Advertising Bureau.

Goldfarb, Avi, and Catherine Tucker (2011), "Online Display Advertising: Targeting and Obtrusiveness," *Marketing Science*, 30 (3), 389–404.

Google (2015), "Where Ads Might Appear in the Display Network," Google Support.

Gordon, Brett, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky (2017), "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," white paper.

Hoban, Paul R., and Randolph E. Bucklin (2015), "Effects of Internet Display Advertising in the Purchase Funnel: Model-Based Insights from a Randomized Field Experiment," *Journal of Marketing Research*, 52 (3), 375–93.

Hu, Ye, Leonard M. Lodish, and Abba M. Krieger (2007), "An Analysis of Real World TV Advertising Tests: A 15-Year Update," *Journal of Advertising Research*, 47 (3), 341–53.

Hunter, Anne, Meredith Jacobsen, Richard Talens, and Tony Winders (2010), "When Money Moves to Digital, Where Should It Go?" comScore white paper, https://www.comscore.com/Insights/Presentations-and-Whitepapers/2010/When-Money-Moves-to-Digital-Where-Should-It-Go.

IAB (2014), "IAB Internet Advertising Revenue Report 2013," http://www.iab.net/AdRevenueReport.

Imbens, Guido W., and Joshua D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62 (2), 467–75.

Imbens, Guido W., and Donald B. Rubin (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press.

Johnson, Garrett A., Randall A. Lewis, and Elmar I. Nubbemeyer (2017), "The Online Display Ad Effectiveness Funnel and Carryover: Lessons from 432 Field Experiments," working paper, https://ssrn.com/abstract=2701578.

Johnson, Garrett A., Randall A. Lewis, and David Reiley (2017), "When Less Is More: Data and Power in Advertising Experiments," *Marketing Science*, 36 (1), 43–53.

Kalyanam, Kirthi, John McAteer, Jonathan Marek, James A. Hodges, and Lifeng Lin (2015), "Cross Channel Effects of Search Engine Advertising on Brick and Mortar Retail Sales: Insights from Multiple Large Scale Field Experiments on Google.com," working paper, https://ssrn.com/abstract=268411.

Lambrecht, Anja, and Catherine Tucker (2013), "When Does Retargeting Work? Information Specificity in Online Advertising," *Journal of Marketing Research*, 50 (5), 561–76.

Lavrakas, Paul J. (2010), "An Evaluation of Methods Used to Assess the Effectiveness of Advertising on the Internet," technical report, Interactive Advertising Bureau.

Lewis, Randall, Justin M. Rao, and David H. Reiley (2015), "Measuring the Effects of Advertising: The Digital Frontier," in *Economic Analysis of the Digital Economy*, Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, eds. Chicago: University of Chicago Press.

Lewis, Randall A. (2014), "Worn-Out or Just Getting Started? The Impact of Frequency in Online Display Advertising," working paper, Google.

Lewis, Randall A., and Dan Nguyen (2015), "Display Advertising's Competitive Spillovers to Consumer Search," *Quantitative Marketing and Economics*, 13 (2), 93–115.

Lewis, Randall A., and Justin M. Rao (2015), "The Unfavorable Economics of Measuring the Returns to Advertising," *Quarterly Journal of Economics*, 130 (4), 1941–73.

Lewis, Randall A., and David H. Reiley (2014), "Online Ads and Offline Sales: Measuring the Effect of Retail Advertising via a Controlled Experiment on Yahoo," *Quantitative Marketing and Economics*, 12 (3), 235–66.

Lodish, Leonard, Magid Abraham, Stuart Kalmenson, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, et al. (1995), "How TV Advertising Works: A Meta-Analysis of 389 Real World Split Cable TV Advertising Experiments," *Journal of Marketing Research*, 32 (2), 125–39.

Manski, Charles F. (2007), *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press

Moriguchi, Takeshi, Guiyang Xiong, and Xueming Luo (2016), "Retargeting Ads for Shopping Cart Recovery: Evidence from Online Field Experiments," working paper, https://ssrn.com/abstract=2847631.

PriceWaterhouseCoopers (2011), "Measuring the Effectiveness of Online Advertising," technical report, IAB France and SRI.

Sahni, Navdeep (2015), "Effect of Temporal Spacing Between Advertising Exposures: Evidence from an Online Field Experiment," *Quantitative Marketing and Economics*, 13 (3), 203–47.

Sahni, Navdeep (2016), "Advertising Spillovers: Evidence from Online Field Experiments and Implications for Returns on Advertising," *Journal of Marketing Research*, 53 (4), 459–78.

Sahni, Navdeep S., and Harikesh Nair (2016), "Does Advertising Serve as a Signal? Evidence from Field Experiments in Mobile Search," working paper, https://ssrn.com/abstract=2721468.

Sahni, Navdeep S., S. Narayanan, and K. Kalyanam (2016), "An Experimental Investigation of the Effects of Retargeted Advertising: The Role of Frequency and Timing," working paper, https://ssrn.com/abstract=2852484.

Shrivastava, A. (2015), "Understanding the Impact of Twitter Ads Through Conversion Lift Reports," Twitter blog.

Simester, D., J. Hu, E. Brynjolfsson, and E. Anderson (2009), "Dynamics of Retail Advertising: Evidence from a Field Experiment," *Economic Inquiry*, 47 (3), 482–99.

Tuchman, A.E. (2016), "Advertising and Demand for Addictive Goods: The Effects of e-Cigarette Advertising," working paper, Northwestern University.

Yildiz, T. and Narayanan, S. (2013), "Star Digital: Assessing the Effectiveness of Display Advertising," *Harvard Business Review* Case Study.