Ghost Ads: Improving the Economics of Measuring Ad Effectiveness

Garrett A. Johnson, Randall A. Lewis & Elmar I. Nubbemeyer*

February 18, 2016

Abstract

To measure the effects of advertising, marketers must know how consumers would behave had they not seen the ads. We develop a methodology we call 'Ghost Ads,' which facilitates this comparison by identifying the control-group counterparts of the exposed consumers in a randomized experiment. We show that, relative to Public Service Announcement (PSA) and Intent-to-Treat A/B tests, 'Ghost Ads' can reduce the cost of experimentation, improve measurement precision, and work with modern ad platforms that optimize ad delivery in real-time. We also describe a variant 'Predicted Ghost Ads' methodology that is compatible with online display advertising platforms; our implementation records more than 100 million predicted ghost ads per day. We demonstrate the methodology with an online retailer's display retargeting campaign, for which a PSA test would be severely biased. We show novel evidence that retargeting can work as the ads lifted website visits by 17% and purchases by 11%. Compared to Intent-to-Treat or PSA experiments, advertisers can measure ad lift just as precisely while spending at least an order of magnitude less.

*Johnson: Simon Business School, University of Rochester, <Garrett.Johnson@Simon.Rochester.edu>. Lewis: Netflix, <randall@econinformatics.com>. Nubbemeyer: Google, <elmarn@google.com>. We thank seminar participants at Kellogg, Stanford GSB, UCSD Rady, Abdelhamid Abdou, David Broockman, Hubert Chen, Mitch Lovett, Preston McAfee, John Pau, David Reiley, Robert Saliba, Kathryn Shih, Robert Snedegar, Hal Varian, Ken Wilbur, and many Google employees and advertisers for contributing to the success of this project.

1 Introduction

Marketers need to know the effectiveness of their advertising. Digitization has advanced this goal by giving marketers unprecedented access to data on clicks, site visits, and online purchases by individual users who see their ads. While granular measurement is now routine in digital advertising, the causal measurement of ad effectiveness lags far behind (Lavrakas, 2010). Randomized controlled experiments can be a simple and effective tool for causal measurement, but current approaches limit their widespread use.

A workhorse approach for advertising effectiveness experiments is the delivery of Public Service Announcements (PSAs) to the control group. PSAs identify baseline purchase behavior among the subset of control users reached by the ads: the experimental ad effect estimator compares outcomes between people who see the advertiser's 'treatment ad' and those who see the PSAs. Unfortunately, PSAs are expensive and prone to errors because they require coordination among advertisers, publishers, and third-party charities. These costs reduce advertiser's incentives to learn through experimentation.

Worse, PSA experiments are rendered invalid when marketers use performance-optimizing computer algorithms to deliver ads. These algorithms use machine learning models to maximize user clicks, site visits or purchases. Performance-optimized campaigns break PSA experiments: to maximize performance of the treatment ad and PSA, the ad platform will assign different types of users to be exposed to the PSA or treatment ad. However, an experiment is predicated on the symmetry of the treatment and control groups: the experimental groups must be the same but for the treatment ad. In this case, the PSA-exposed users are no longer a valid holdout group for the treatment-ad-exposed users. This failure of PSAs is central for the online display ad industry because two thirds of dollars spent there make use of performance-optimizing technology (IAB, 2014).

We propose a new methodology for ad effectiveness field experiments: *Ghost Ad*s. Like PSAs, 'ghost ads' identify ads in the control group which *would have been* the focal advertisers' ads had the user been in the treatment group. Unlike PSAs, ghost ads are compatible

with performance-optimizing technology and avoid the costs of PSAs. Instead of PSA ads, the control group user sees whatever ads the ad platform chooses to deliver when the treatment ad is absent. The Ghost Ad methodology is implemented at the ad platform level. Normally, the ad platform runs an auction to determine which ad it will show to a user, then logs these ads in a database. To determine when the focal ad would have been shown to a user 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 logs *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 users. The Ghost Ad method can be applied across several digital media—e.g. search, online display (including video and radio), addressable television—whenever consumers can be randomly sorted into treatment groups.

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 the attention of consumers. When the focal advertiser exits the marketplace, many other advertisers take its place including some direct competitors—creating an externality on the focal advertiser. For instance, advertising can play 'defensive' role that blocks competitors from stealing away consumers in addition to advertising's 'offensive' role that pulls in consumers. Since PSAs are chosen to be orthogonal to the advertiser, PSA estimates ignore the strategic consequences of the competing ads they displace. If advertising has a defensive role on net, 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 in cases where the ad platform and a downstream firm jointly control ad delivery. For instance, many online display ad platforms do not fully control ad delivery because the platform merely suggests 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 here misses the

unobservable cases in the control group where the platform would suggest the focal ad but the publisher would refuse it. This slippage breaks the symmetry between the focal ads in the treatment group and the ghost ads in the control group. The Predicted Ghost Ad methodology preserves this symmetry by running a simulated auction across *both* treatment and control group users 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 a another database to flag occasions where the ad platform intends to serve an experimental ad regardless of the user's actual treatment assignment. Effectively, the predicted ghost ads help construct a powerful instrumental variable in order to calculate the Local Average Treatment Effect (LATE, see Angrist et al. 1996) among users who are predicted to be exposed to the advertising campaign. We implement the Predicted Ghost Ad methodology successfully with Google's online display network, which we demonstrate using an early application. Now, Predicted Ghosts Ads are broadly used with over 100 million predicted impressions daily. In follow-up work, we describe the results of over 400 advertiser field experiments averaging over 5 million users each over a four month period (Authors, 2015).

We apply our Predicted Ghost Ad methodology to show novel evidence that an online display retargeting campaign generates incremental conversions. The effectiveness of retargeting ads is controversial. While consumers who see retargeted ads may have high sales, this association may not be causal because the exposed users are a highly selected group that may have purchased anyways. Retargeting could even reduce ad effectiveness if it provokes reactance in consumers. While Lambrecht & Tucker (2013) and Bleier & Eisenbeiss (2015) compare the relative performance of different retargeting creatives, we provide the first evidence that retargeting works when compared with a control. We work with an online sports and outdoors retailer to retarget users who visited product pages on its website. We find that the retargeting campaign increases website visitors by 26.6% (t=40.38), transactions by 12.0% (t=5.42), and sales by 10.8% (t=3.45). This retargeting campaign was performance-

optimized and incompatible with PSAs, but our Predicted Ghost Ad methodology succeeds and demonstrates highly significant lifts.

The paper is organized as follows. Section 2 outlines the design of an ad experiment and explains the existing PSA and Intent-to-Treat approaches. Section 3 explains the Ghost Ad methodology and explains how Ghost Ads and PSAs perform as ad platforms have evolved. Section 4 describes the related Predicted Ghost Ad methodology. In Section 5, we describe our empirical application and provide evidence that a retargeting campaign can increase sales. Section 6 outlines some challenges in implementing Ghost Ads. Section 7 concludes.

1.1 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—since this cost makes advertisers reluctant to use PSA tests and willing to sacrifice measurement precision for smaller control groups. Of the 25 online display experiments at Yahoo! listed by Lewis & Rao (2015), only 8 use PSAs and the average control group allocation is only a third of users though an even split maximizes precision. In Hoban & Bucklin (2015), the advertiser wanted less than 2% of users in the control group. 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 and possibly promote its effectiveness as in Sahni (2015) or Johnson et al. (2014). Second, advertisers may engage research firms to design the experiment as in the brand survey studies by Goldfarb & Tucker (2011a,b) and Bart et al. (2014). Third, Lewis (2010) and Lewis & Nguyen (2015) use natural experiments where the ad platform randomly split the delivery of unrelated ad campaigns.

Intent-to-Treat A/B[1] experiments also eliminate the costs of PSAs, but lack the measurement precision of ghost ads or PSAs. Without PSAs, we do not observe the control

[1]Intent-to-Treat experiments are similar to A/B or bucket tests in that A/B testing platforms typically use the Intent-to-Treat estimator when analyzing experiments. However, we distinguish Intent-to-Treat experiments from classical A/B tests in order to emphasize that treatment take-up is endogenous in our setting–not all users are treated. See Section 2.

group users who would be exposed to an ad when the campaign's reach is incomplete. As we see in Section 2.1, Intent-to-Treat experiments compare all treatment and control eligible users including unexposed users who contribute only noise to the estimator. As such, this approach requires high advertising expenditure as treatments to detect significant effects like in the search ad experiments by Blake et al. (2013) and Kalyanam et al. (2015). As in those studies, Intent-to-Treat is compatible with geographic- rather than user-level randomization when the latter is infeasible. High spend does not guarantee significant Intent-to-Treat results though, as in the study of online display ads and retail sales by Lewis & Reiley (2014).

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* that vary the quantity of ads or *copy tests* that vary creative content. The majority of the split-cable TV advertising studies in Lodish et al. (1995) and Hu et al. (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 & Nair, 2015). Nonetheless, both copy tests and PSA tests are biased when they employ performance-optimizing technology, which researchers must avoid until now. The Ghost Ad methodology makes ad experiments compatible with these technologies used broadly in Internet advertising.

To date, retargeting studies use copy tests to evaluate the personalization of retargeting creatives. Lambrecht & Tucker (2013) compare dynamic retargeting ads that include the product with which the consumer engaged (the 'focal product') to generic brand creatives. Bleier & Eisenbeiss (2015) compare retargeting ads that feature products from the focal product's category and/or manufacturer's brand to randomly-selected products. Lambrecht & Tucker (2013) find that personalized ads reduce transactions overall whereas Bleier & Eisenbeiss (2015) find the opposite for click-through rates. Both studies agree that content targeting is important for personalized retargeting ads whose effectiveness improves when

the user browses content related to the focal product (Lambrecht & Tucker, 2013) or browses shopping sites (Bleier & Eisenbeiss, 2015). Here, we instead wish to evaluate the effectiveness retargeting against a holdout group to evaluate whether retargeting reaches consumers who would purchase anyway or causes incremental purchases. Our application uses dynamic retargeted advertising like in Lambrecht & Tucker (2013) and is the first retargeting study to examine effects on sales.

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

2 Experimental design & existing approaches

In this section, we introduce the experiment's design and nomenclature that are the foundation of our Ghost Ad methodology. We focus here on the online display advertising context, though much of the discussion applies to other ad media. We then discuss the strengths and limitations of existing ad experiment approaches in two subsections.

An online advertising experiment involves *advertisers*, *users*, and an *ad platform*. Users are represented by a unique identifier such as a web browser cookie. The ad platform matches advertisers to the ad opportunities or *impressions* that users generate when browsing webpages.

Suppose we want to estimate the effectiveness of an online display ad campaign for Christian Louboutin, a French manufacturer of high-end ladies shoes with distinctive red

soles. Typical advertiser objectives include generating reach, brand favorability, clicks, and conversions. *Conversions* are user interactions with the advertiser such as visits to the advertiser's website, sign-ups, or purchases. Suppose that Louboutin wants to maximize visits to its website.

To assess the Louboutin campaign's effectiveness, we use an experiment that will compare the visits to Louboutin's website by users in a Louboutin treatment group with visits by users in a holdout or control group. The experiment randomly assigns each user to either the *treatment* group that is eligible to see the Louboutin ad or the *control* group that is ineligible. The experimental randomization ensures the two groups are equivalent, so that we can attribute any changes in outcomes to the Louboutin ad. The ad platform's random assignment rule may operate on a previously selected list of eligible users. The ad platform may instead determine eligibility and treatment assignment 'on the fly'—meaning in real time as users arrive at eligible websites. A user in the treatment group is *treated* if he is exposed to the *treatment ad*.

Whether a treatment user is treated depends on both the ad platform's ad delivery decisions and the user's browsing decisions. Sometimes, the platform will reserve certain ad impressions exclusively for the experiment (e.g., see Lewis & Nguyen 2015). Often, the platform must choose whether to display the treatment ad or another ad at each opportunity. While a user does not explicitly choose whether to see a particular ad, the user implicitly affects treatment take-up through her browsing: choosing when, which websites, and how many pages to visit. That is, the user determines the number of opportunities to receive the experimental ad. However, the user cannot decide to specifically avoid the experimental ad sent by the ad platform.[2]

Figure 1 illustrates the design of an ideal ad experiment. Figure 1 categorizes users by whether they belong to the treatment or control group and whether the users are treated

[2]Users may use ad blocking technology to avoid all ads, but this is orthogonal to treatment assignment. Some companies such as Google and Facebook allow users to 'X-out' and avoid specific ads, though X-out rates are negligible.

or not. We represent consumer heterogeneity in purchasing behavior by two types of users: those who wear striped shirts and those who wear solid shirts. Here, the advertiser is targeting the striped shirt users who form a majority among the treated. Among control group users, we distinguish between the *counterfactual treated* users who *would* have been exposed to the treatment ad and those who would not. Consequently, the distribution of types in Figure 1 across treated and untreated is identical across treatment groups, which illustrates the experiment's symmetry. In this ideal setting, we can evaluate the ad effect with the *Treatment on the Treated* (TOT) estimator, which differences the outcomes of the treated and counterfactual treated. However, in practice, while we can always split the treatment groups into users who have and have not seen the Louboutin ad, we do not observe the equivalent subgroups within the control group. Below, we discuss two dominant approaches to this problem: Intent-to-Treat and Treatment-on-the-Treated using PSAs.

2.1 Intent-to-Treat approach

If we do not observe the counterfactual treated users, we can still evaluate the ad effect with the *Intent-to-Treat* (ITT) approach. ITT is a valid experimental approach whose estimator differences the outcomes of the entire treatment group and the entire control group. Figure 2 illustrates the logic of ITT where the counterfactual treated users from Figure 1 are no longer distinguishable. The ITT approach does not require PSAs and is easy to implement.

The drawback of this approach is that the ITT estimator is less precise than the TOT when the variances in outcomes are comparable between the treated and untreated. To see this, we split the ITT estimator into two components: 1) the experimental difference among the treated (TOT) and 2) the experimental difference among the untreated. Logically, the latter difference should be zero, but the empirical difference contributes noise that weakens precision.

ITT's precision is poor when the proportion of untreated users 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 1. If all impressions are eligible, the experimenter can only restrict the set of users to those who were online to see an ad during the campaign. In this case, the proportion of treated users can be very small (e.g., 3% or less). The problem improves if the experimenter has a pre-defined list of eligible users in the experiment.[3] However, ad platforms often determine eligibility 'on the fly' and therefore cannot generate such a list of eligible users, leaving the set of all online users as the only valid option. Given that ITT reduces precision, we would like to identify the counterfactual treated users and estimate TOT.

2.2 PSA approach

The other dominant approach is to use *Public Service Announcements (PSAs)* in the control group to identify the counterfactual treated users. The ad platform must deliver the PSAs to the control group in the same way as it delivers the treatment ads to the treatment group. Then, we can achieve the ideal case in Figure 1 where we can compare outcomes among the treated between treatment and control groups. In the experimental context, PSAs are often charity ads (see e.g. Hoban & Bucklin 2015; Yildiz & Narayanan. 2013) but they are more generally neutral ads with an orthogonal call to action to the treatment ad. For our purposes, the PSA approach is equivalent to showing control users an ad from an unrelated company (see e.g. Lewis 2010; Lewis & Nguyen 2015), a blank ad (see e.g. Goldfarb & Tucker 2011a,b; Bart et al. 2014), or a 'house ad' advertising the publisher (see e.g. Johnson et al. 2014; Sahni 2015). In our example, suppose the PSA advertiser is a nonprofit sea turtle

[3]Some advertisers perform their own randomization of users into separate treatment and control group targeting lists for Intent-to-Treat experiments. However, recently developed 'similar audiences,' 'look-alike,' and 'broad match' technologies seek to increase the number of eligible users beyond such pre-defined lists by identifying users with similar characteristics. These technologies can lead to control-group users being exposed to the treatment ads if the list of eligible users is increased after (or independent of) the advertiser's randomization. While such experiments can still be analyzed using the local average treatment effect (LATE) estimator (Angrist et al., 1996) to deal with these 'always-takers' of the ad treatment, the straightforward interpretation of the experiment's results and their generalizability to the entire campaign are jeopardized. In short, the ad expenditures on the always-takers in both treatment and control at best do not help or at worst add noise to all experimental measurements for the uncontaminated compliers. That said, our predicted ghost ads implementation described in Section 4 should be immune to this problem by performing the randomization after all eligibility information has been considered.

rescue organization. Hence, an *experimental ad* refers to either a *treatment ad* (Louboutin) or *control ad* (sea turtle).

The main advantage of PSAs is that they increase the precision of the experimental estimates in two ways. First, relative to ITT, PSAs allow the experimenter to prune away users who do not see an ad and contribute only noise to the estimator. Second, PSAs allow the experimenter to prune away outcomes prior to the first ad exposure. Since the outcome cannot be affected before the initial ad exposure, this component of the outcome variable also only adds noise. This approach is valid when the first experimental ad is delivered symmetrically across treatment groups. As an example, Johnson et al. (2014) show that PSAs can significantly improve precision: their TOT estimate is 31% more precise than their ITT estimate, which is equivalent to a 110% increase in sample size. Post-exposure filtering accounts for a quarter of this improvement in precision.

The first drawback of PSAs is their cost. The ad platform and the experimental 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 advertisers must negotiate splitting these costs and coordinate to obtain a suitable PSA. Additionally, the ad platform must be configured to handle the PSAs in exactly the same way as the treatment ad. Small errors are expensive because they can invalidate the experiment by introducing selection bias or complicate any analysis that salvages the experiment. Finally, the cost of PSAs does not fall with scale, making PSAs an obstacle to large-scale, automated experimentation.

The second drawback of PSAs is that modern ad platforms deliver a PSA campaign and the treatment campaign differently so that PSAs do not serve as valid control ads. This phenomenon is depicted in Figure 3, where we see that the PSA reaches the same number of users but reaches a different distribution of user types. Hence, any measured difference in sales between exposed users in the treatment and control groups will conflate the causal ad effect with the user-type selection bias. This occurs because modern ad platforms optimize ad delivery by matching each ad to different user types. We discuss this limitation in Section 3.

3 Ghost Ad methodology

In this section, we explain the Ghost Ad methodology as an inexpensive alternative to using PSAs as control ads. We then describe how, as ad platforms become more sophisticated, the Ghost Ad methodology succeeds where the PSA methodology fails.

With the Ghost Ad methodology, the ad platform delivers ads to the control group normally, but tags *ghost ad impressions* where the ad platform *would* have served an experimental ad. We name our solution 'ghost ads' because the ghost ad tags are invisible to the user as if the ghost ad 'possesses' the ad that is actually shown. Ghost ads avoid the ad inventory and coordination costs of PSAs. Ghost ads also reduce the scope for error in planning and executing experiments. By eliminating the cost of PSAs, ghost ads facilitate widespread experimentation.

Figure 4 illustrates how the Ghost Ad methodology works. Figure 4 shows six ads delivered to an identical user in four settings: 1) treatment group, 2) control group without control ads, 3) control group with PSA ads, and 4) control group with ghost ads. In the first setting, the treatment group user sees three Louboutin shoe ads and three ads from other advertisers. In the second setting, the control group user see three ads from other advertisers instead of three Louboutin ads. In the third setting, the PSA group user sees three sea turtle rescue charity ads in place of three Louboutin ads. Finally, in the fourth setting, the ghost ad user sees the same three ads as in the second setting without control ads, but the ad platform knows which three ads are tagged by Louboutin's ghost ads—depicted by overlaid ghosts. In this example, the ad platform delivers identical quantity and order of experimental ads in the treatment, PSA, and ghost ad settings.

The Ghost Ad methodology is needed to perform experiments on modern ad platforms that match ads to users in ways that invalidate the use of PSAs for experiments. To demonstrate the limits of the PSA and ghost ad technologies, we describe their performance as ad platforms have evolved. Toward this, we imagine the evolution of ad platform technology as divided into three generations:

1. First generation: *Reach-based ad platforms* allocate ads according to an advertiser's demographic and content targeting criteria in order to deliver the agreed-upon number of impressions during the campaign period. By ignoring campaign performance feedback, both PSAs and Ghost Ads work.

2. Second generation: *Action-optimized ad platforms* serve ads to maximize a given action like a campaign's click or conversion performance. To do so, action-optimized systems employ *aggregate* feedback from the campaign's performance across websites and ad creatives by user characteristics. This feedback causes PSAs to fail, but Ghost Ads still succeed.

3. Third generation: *User-optimized ad platforms* further optimize campaign performance delivery to individual users based on that user's click and conversion activity. The *user-level* feedback invalidates PSAs, but Ghost Ads still identify the counterfactual treated users because the first experimental ad delivered to each user has no user-level feedback.

We can roughly equate the different ad platform generations to different ways an advertiser can purchase ads. An advertiser who wants to buy a certain quantity of impressions paid for on a cost per mille/thousand (CPM) basis only needs reach-based ad platform technology. An advertiser who instead optimizes for clicks or conversions buys on a cost-per-click (CPC) or cost-per-action (CPA) basis and uses the capability of an action-optimized ad platform. An advertiser who wishes to target individual users who have interacted with the advertiser online—called *retargeting*—requires a user-optimized ad platform.

In the next three subsections, we contrast the performance of PSAs and ghost ads as control ads in each of the three generations. We show that PSAs are invalid after the first generation, whereas ghost ads identify the counterfactual treated users throughout. Table 1 serves as a reference for the reader by outlining the performance of the two approaches across all three generations.

3.1 First generation: Reach-based ad platform

A reach-based ad platform delivers the treatment, PSA, and ghost ads identically, which preserves the symmetry of ad delivery across experimental groups. A reach-based platform encompasses everything from TV and radio ad schedules to the impression-based delivery that emerged in Internet advertising. A sophisticated reach-based ad platform allocates ads using rich advertiser, site, user, and time-varying inventory variables. Crucially, the reach-based ad platform ignores feedback from the users' interactions with the ads or subsequent behavior like visiting the advertisers' site. Hence, a reach-based ad platform treats a treatment ad, PSA, or ghost ad symmetrically provided they are configured the same way within the ad platform.[4] Such configuration parameters include ad quantity, time period, user-type, and webpage targeting attributes, and budget. Thus, by construction, PSAs must cost the same as the treatment ads, whereas ghost ads are free.

Ghost ads deliver the correct baseline behavior because users see the ads that they *would* have seen without the experiment. The 'possessed' ad impressions shown to the control group are for many advertisers rather than concentrated on PSAs for a single charity. Further, if the counterfactual ads include those from the advertiser's competitors, these could capture business from the advertiser. This is Louboutin's relevant competitive baseline to evaluate what happens if it does not advertise, not the artificial comparison with a sea turtle rescue PSA.

If treatment ads affect user behavior, as intended, these treatment effects can alter subsequent ad exposures, but such differences are usually small enough to ignore. For example, a treated user may click on the Louboutin ad and leave the publisher's site to browse Louboutin's online storefront. As a result, the user could have fewer opportunities to see ads

[4]Some ad platform require the treatment and PSA campaigns in a PSA test to be pre-configured prior to the start and disallow any changes during the experiment. This constraint is imposed because even two symmetric changes made simultaneously to two active campaigns through a campaign management user interface can have large asymmetric propagation delays to each of the individual ad serving machines, leading to asymmetric ad serving behavior until all of the ad servers recognize the change. For example, if the advertiser changes their CPM bid or targeting criteria for the experiment, one of the two campaigns (treatment or PSA) might fully deploy an hour earlier than the other, leading to divergent ad serving.

than her counterpart in the control group who was not offered the opportunity to click on (or see) the Louboutin ad. That is, user take-up of subsequent experimental ads (treatment intensity) is *endogenous*. In practice, we have found these differences to be negligible. For instance, Johnson et al. (2014) do not detect any significant difference in experimental ad exposures between treatment groups. In any case, while this difficulty is fundamental to any experimental methodology that seeks to identify counterfactual impressions, this treatment effect feedback should not affect the delivery of the first experimental impression.

3.2 Second generation: Action-optimized ad platform

Action-optimized ad platforms match ads to users in order to optimize each campaign's performance. This is most commonly done by allowing advertisers to target their campaigns by purchasing ads on a CPC or CPA basis. To implement this, online display ad platforms run an auction for each ad impression where CPC bids are weighted by the predicted probability that the user will click on the advertiser's ad. The click prediction is based on a complicated combination of the advertiser's and user's historical click rates and those of users with common characteristics (e.g. demographics). Such a system seeks to maximize clicks by controlling *who* sees the ad, *how many* ads they see, and both *where* and *when* the ad is shown. For this reason, we say that the action-optimized ad platform's ad serving is *endogenous*: a user's behavior affects treatment intensity of those who look and behave like her. We might imagine that the ad platform would want to change ad serving at the individual-level based on whether the user clicks and or converts. This in-campaign individual optimization further complicates the ad platform's operation, so we defer this discussion to the next subsection. For now, we assume that the ad platform only optimizes using ad, advertiser, site, user characteristics, time, and *pre-campaign* user-level behaviors.

The action-optimized ad platform biases ad delivery so that the PSAs are no longer valid control ads. Consider our example Louboutin campaign with sea turtle rescue PSAs. People who are interested in Louboutin shoes are likely quite different from those interested

in rescuing sea turtles. The action-optimized ad platform uses machine learning to "learn" these preferences from click behavior and deliver the ads to the types of users who click on each creative. This creative-level optimization means that the ad delivery of *any* two distinct creatives, even with the same targeting configuration, will differ. Hence, ad delivery would be unbalanced between the Louboutin ad and PSA.[5] 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 ad users and the turtle-rescue users will be biased. By the same logic, we get biased estimates for attributing ad effects to different websites or user demographics when using PSAs as control ads.

Figures 3 and 5 illustrate how PSA campaigns bias the delivery of experimental ads. Recall from Section 2 that Figure 3 shows how the distribution of treated user types will vary between the treated and the PSA treated; whereas, the Ghost Ad methodology will balance these types as in Figure 1. Figure 5 compares the timing and quantity of ads delivered to a user under treatment ads, ghost ads, and PSAs. While the ad serving decisions under treatment ads are the same under ghost ads, the PSA user sees fewer experimental ads at different times than the treatment user in the action-optimized platform. Hence, the technique from Section 2.2—filtering out pre-exposure outcomes to improve measurement precision—can be done symmetrically with treatment ads and ghost ads but not PSAs in a second-generation ad platform.

[5]Google's "DCM Audience Segmentation" product relies on PSAs but implements a clever trick to avoid this pitfall of unbalanced ad serving. This system builds a single campaign with the two experimental ads: the Louboutin and sea turtle ads. The system creates treatment and control groups by randomizing which users exclusively see one or the other ad. Then, Google uses third-party ad serving to trick the optimization platform into thinking that the two creatives are really the same, preventing it from treating them differently. While this system ensures symmetric delivery, the experimental estimates it delivers would reflect the effect of ads on users whom the ad platform believes are shoe-lovers and nature-lovers rather than just shoe-lovers. Furthermore, this solution has all the other costs of PSAs.

3.3 Third generation: User-optimized ad platform

Finally, we examine the user-optimized ad platform, which has all the characteristics of an action-optimized ad platform plus the ability to customize user-level ad serving during a campaign based on each user's actions. Since user-optimized ad platforms build on action-optimized platforms, we know that PSAs will not work. The capability to tailor delivery to the user enables in-campaign retargeting: if the consumer interacts with the advertiser's website or advertisement, the ad platform will infer the user's interest and alter ad serving. Often, the ad platform will show interested users more of these ads, but the platform may reduce the number of ads, say, after a purchase. Again, the user-optimized platform's ad serving is *endogenous*: the user's behavior affects her treatment intensity. In-campaign retargeting capability is valuable to ad platforms because the value of a match between the user and the advertiser is largely governed by the user's time-varying taste for the advertiser (e.g., whether user is in-market), not her pre-exposure observable characteristics. As we will see, the distinction between action-optimized and user-optimized ad platforms is that the Ghost Ad methodology no longer mirrors the treatment group's ad delivery after the first impression.

The combination of endogenous ad serving and endogenous user take-up alters the ghost ad delivery. In Section 3.1, we describe the endogenous take-up problem where users alter their behavior in response to the treatment effect and thereby alter their opportunity to receive subsequent impressions. While we could ignore this with the reach-based ad platform, we cannot ignore this problem with the user-optimized ad platform. The user-optimized ad platform mechanically magnifies the endogenous treatment problem because the incremental users driven to action by the ads will receive dramatically different quantities of subsequent impressions. This difference is fundamental to third-generation platforms with endogenous delivery: any implementation of the Ghost Ad methodology cannot know which users would have responded to the treatment ad to correspondingly ramp up ad delivery. This resembles the impossibility of identifying the control group's counterfactual treatment-ad clickers: we

can never know who in the control group would have clicked on an ad without actually showing the ad.

Figure 5 illustrates the divergence of the Ghost Ad methodology in third-generation ad platforms. Midway through the campaign, the Louboutin ads *cause* the user to visit the Louboutin website when she is in the treatment group. The site visit is causal because she would not have visited the Louboutin site under ghost ads or PSAs. After the Louboutin site visit, Figure 5 shows that the third generation platform serves more impressions when she is in the treatment group. The Ghost Ad approach performs identically until the causal visit but then diverges afterwards by delivering fewer ads. Of course, ad delivery under PSAs in Figure 5 diverges from the treatment group throughout the campaign as in the action-optimized case.

The Ghost Ad methodology no longer exactly replicates the treatment group's ad delivery but still delivers unbiased treatment effect estimates. Most importantly, the Ghost Ad methodology performs exactly as the treatment group until the ads cause the behavior of the treatment group user to diverge because the third-generation platform's user-level information set is equivalent across the treatment and control group initially. Hence, each user's first experimental ad will be identical regardless of treatment or control assignment because the ad server has not yet treated or withheld any treatment based on that user's assignment. The first ad suffices to unbiasedly estimate the experimental differences in advertiser outcomes by identifying the counterfactual treated users in the control group. The identical timing allows us to ignore pre-treatment outcomes in the treatment and control groups which improves the precision of the TOT estimator. However, since the distribution of experimental ad impressions will differ between the treatment and control groups, a simple comparison of treatment users with control users having the same frequency of experimental ads will be biased.

4 Predicted Ghost Ad methodology

In theory, the Ghost Ad approach is a fitting solution for ad effectiveness experiments. In practice, the technology is difficult to implement with current Internet ad platforms, because they were not built with ghost ads in mind. Instead, we propose a *Predicted Ghost Ad* methodology as a more robust alternative. The Predicted Ghost Ad methodology predicts rather than determines whether the platform will serve an experimental ad. We use the Predicted Ghost Ad methodology to compute the Local Average Treatment Effect of the ads among *predicted–exposed* users. Below, we describe the challenges in implementing the Ghost Ad methodology in the case of online display ad platforms and outline the Predicted Ghost Ad alternative.

The challenge of implementing the Ghost Ad approach centers on the fact that display ad platforms do not control the full ad delivery process. Display ad platforms are not just a marketplace where advertisers and publishers meet, but rather a network of interlocking demand- and supply-side platforms with imperfect information and incomplete integration (see e.g., Evans 2009 and Johnson 2013). Above all, ad platforms must allocate each ad impression while the webpage loads within a tenth of a second from receiving the ad request. When an ad platform sends an impression to a publisher, the impression can be rejected for a host of reasons. The publisher may have a secret reserve price that the advertiser's bid does not beat. The publisher may also have hidden exclusions that block certain advertisers. 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 ad platform will backfill the impression with another ad. If an ad platform does not control the full ad delivery process, then the platform cannot tag all the counterfactual impressions symmetrically. Perhaps a display ad platform built with the Ghost Ad methodology from the ground up could resolve this challenge, but existing display ad platforms require an alternative approach.

We propose the Predicted Ghost Ad methodology that can work with existing ad platforms. Since a display ad platform does not control which ads ultimately render, Predicted

Ghost Ads try to predict whether an experimental ad renders using a two-stage auction. The Predicted Ghost Ad innovation lies in the first-round simulated auction[6] in which the treatment ad participates for both treatment and control users to generate a symmetric prediction of experimental ad exposure.[7] A *predicted ghost ad impression* then is an ad opportunity where the treatment ad wins the simulated auction, regardless of the user's actual treatment assignment. The second-round real auction then selects ads for the treatment and control users to be sent to the publisher, excluding the treatment ad from the control group's auction. Thus, whereas ghost ads only tag control users, predicted ghost ads tag *both* treatment and control users. Figure 6 illustrates how the Predicted Ghost Ad methodology treats control and treatment users in both the simulated and real auctions. Using predicted ghost ads, we compute the Local Average Treatment Estimate (LATE) for the ad lift among users who were predicted to see at least one experimental ad. Denoting indicator variables for Treatment as T, predicted ghost ad exposure as PGA and actual exposure to the experimental ad as X, we write our PGA LATE estimator for outcome y as

$$PGA\ LATE = \frac{E\left[y|PGA=1,T=1\right] - E\left[y|PGA=1,T=0\right]}{\Pr\left[X=1|PGA=1,T=1\right]} \tag{1}$$

In words, the PGA LATE estimator rescales the experimental difference among predicted exposed users by the probability a user receives a treatment ad conditional on predicted exposure.[8] In our empirical application (Section 5), this probability is 99.9%, which means the simulator rarely over-predicts treatment.

The interpretation of PGA LATE hinges on whether the Predicted Ghost Ad methodol-

[6]Ideally, the Predicted Ghost Ad methodology simulates using the ad server's real auction or ad selection algorithm in order to improve the accuracy of the simulation. However, the requirement is simply that the simulator has sufficient predictive power.

[7]A frequent misconception is to compare the exposed users in the treatment group to the predicted exposed users in the control group. If our Predicted Ghost Ad implementation is imperfect in its predictions, such a comparison would introduce selection bias. For example, if we over-predict ad exposure by 10% of users in a balanced experiment, we could end up comparing 1,000 treatment users to 1,100 control users, a difference that immediately raises doubts in the analysis of the experiment.

[8]While our implementation uses a binary prediction $PGA = \{0, 1\}$, PGA could instead employ continuous probability estimates with $PGA \in [0, 1]$ and estimate the LATE with an instrumental variable regression.

ogy can *under*-predict the experimental ad exposures. That is, whether a treatment ad can ever win the real auction but lose the simulated auction. If no underprediction occurs, then the PGA LATE estimator captures the effect of the ads on *all* exposed users and provides an unbiased estimate of the average effect of treatment on the treated (TOT). If however underprediction occurs, then there exist users who are exposed to the experimental ad but not predicted to be exposed. In our empirical application, underprediction is small: only 3.2% of users are treated but not predicted to be treated. These underpredicted users see an average of 1.5 ads whereas predicted-exposed users see an average of 20.2 ads. Since we predict 96.8% of exposed users and predicted-exposed users see 99.8% of the campaign's ads, we expect our PGA-LATE estimate to well approximate TOT.[9]

In our beta implementation of the Predicted Ghost Ad methodology, we have seen this under-prediction problem crop up due to overlapping predicted ghost ad campaigns. For instance, if users are in overlapping ghost ad experiments for Christian Louboutin and IKEA, then IKEA could win the simulated auction so that an IKEA predicted ghost ad is logged but not one for Louboutin. However, if the user is in the IKEA control group and the Louboutin treatment group, the actual auction excludes IKEA and the Louboutin ad could win. Thus, our Predicted Ghost Ad system under-predicts the Louboutin ad.

Our solution here is to either augment the simulated auction with the 'auction isolation' approach or to re-analyze the experiment with more general 'ghost events.' To implement the *auction isolation* approach, we run a separate simulated auction in the first stage for each overlapping Predicted Ghost Ad campaigns to eliminate the cross-campaign externalities. Auction isolation works well because it maintains the full statistical power of PGA LATE. Absent this, we can fall back on expanding our notion to a *ghost event* which track the users' counterfactual exposure to the ad changes. A ghost event is like 'exposure logging' in Bakshy

[9]The under-prediction problem can cause situations where users are exposed to a treatment 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 Section 3. Though a treatment ad exposure may distort subsequent predicted ghost ad impressions, the underprediction problem so small here that we do not expect or detect an impact on outcomes prior to predicted treatment.

et al. (2014) in that ghost events log activities that are necessary (but perhaps not sufficient) conditions for experimental exposure. For instance, *'all predicted ghost ads'* is a ghost event we can use to compare treatment groups among the users who get a predicted ghost ad for *any* experiment in the system. This solution avoids the cross-experiment externality problem and maintains symmetry between treatment groups, though the corresponding 'All-PGA LATE' estimator is less powerful because it includes more unexposed users. Other examples of ghost events include *ghost bid*s, when the advertiser bids in any of the user's auctions, and *ghost cookied*, all users for whom the advertiser has a cookie.

5 Empirical application

Our empirical application features an advertising experiment for an online retailer of apparel and sporting goods. A confidentially agreement prevents us from naming the advertiser which we will call 'Sportsing Inc.' for the sake of exposition. Sportsing ran a retargeting campaign optimized on consumer transactions that uses the modern third-generation ad platform's capabilities. This setting is thus apt for ghost ads as we have seen that PSAs would deliver invalid ad effectiveness estimates. We demonstrate that our Predicted Ghost Ad implementation delivers the predicted treatment impressions symmetrically across treatment groups. Additionally, we show original and strong evidence that a retargeting campaign lifts both website visits and purchases at Sportsing.

5.1 Experimental design

Sportsing ran a retargeting display advertising campaign in winter 2014.[10] The experiment assigned 70% of users to the treatment group. The campaign delivered 9 million ad impressions to desktop platforms during the two-week experiment. In contrast to previous research, the campaign ran continuously before and after our two-week experiment, and our

[10]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 to running tests.

Predicted Ghost Ad implementation enabled us to turn on the experiment by splitting users into treatment groups at our convenience. Sportsing ran the retargeting campaign on the Google Display Network of over 2 million websites, including Google AdSense publishers (Google, 2015). By retargeting campaign, we mean that Sportsing targeted users who visited its website within the past 30 days—updated on a rolling basis. In particular, eligible users must have either 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 no longer eligible to see the experimental ads. Sportsing optimized the campaign on a cost per conversion basis, where Sportsing fixed the conversion to be a purchase among users who clicked on the ad. 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 conversion and remarketing pixels.

For the campaign, Sportsing employed dynamic retargeting display ads which feature the products that the individual user had viewed on the advertiser's site. Figure 7 shows a similar Google Dynamic Remarketing creative example. 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 smaller Google AdWords creatives similar to search ads—a hyperlink at the top followed by short lines of text—but with a small picture of a single product on the left. In all cases, the ads do not promote a sale.

5.2 Experimental validation

Experiments must demonstrate the validity of the randomization by showing that the treatment groups are equivalent before they receive the treatment. To do so, we test for differences in the subjects' characteristics and pre-treatment outcomes. These checks are crucial here because they also validate that our Predicted Ghost Ad implementation performs as desired. To further validate our implementation, we also test for differences in the delivery of the first predicted impression across treatment groups. Table 2 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=0.34, 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 2, we compare the distribution of two categorical variables that indicate the user's location: country and city. Chi-squared tests fail to reject (p=0.85 and p=0.50).

Next in Table 2, we test the predicted ad exposure variables for users' first predicted impression. As we discuss in Section 3, a retargeting campaign will *only* replicate the delivery of the *first* predicted ad across the treatment and control groups. Hence, 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 most common websites in the campaign (p=0.49). Moreover, both groups see the same distribution of the first ad's creative format: flash or text formats as well as different ad shapes and sizes (p=0.68). We also find no significant differences for t-statistic

tests of equality of means for the first predicted ad's cost per click (p=0.38), predicted conversion probability (p=0.10), predicted click probability (p=0.47), and predicted cost per impression (p=0.10). The ad platform calculates these variables for each impression as they affect the platform's delivery decisions. 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 pre-campaign outcomes. In Table 2, we see no differences in site visits (p=0.92), transactions (p=0.76), and sales (p=0.75) 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 since we want to filter out this data to increase the precision of our estimates. Again, Table 2 shows no significant differences (p=0.33, 0.21, and 0.11), instilling confidence that our experimental estimates represent the ad lift and not a failure of our Predicted Ghost Ad implementation.

While the first predicted ghost ad should be symmetric across the treatment and control groups, subsequent impressions should not. As discussed in Section 3, 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. As expected, Table 2 reveals significant differences across treatment groups in variables that average over *all* predicted impressions per user. Indeed, a lack of difference between subsequent treatment and control impressions might be symptomatic of an ineffective (retargeting) campaign—one which fails to produce the mechanical feedback of a user-optimized ad platform.

5.3 Results

Now we apply our Predicted Ghost Ad methodology to measure the effect of Sportsing's retargeting ads on sales and site visits. Recall from Section 4 that the Predicted Ghost Ad methodology runs a first-stage simulated auction for both treatment- and control-group users to predict whether Sportsing is expected to win the second-stage real auction when participating (in the treatment group). We use the Predicted Ghost Ad system to estimate the Local Average Treatment Effect (LATE) from equation (1) on the predicted exposed users. Table 3 lists both the Predicted Ghost Ad system's LATE estimates as well as the Intent-to-Treat (ITT) estimates that make no use of the Ghost Ad system.

Our Predicted Ghost Ad system's LATE estimates yield highly significant evidence that Sportsing's ads increase both site visits and sales. Following Johnson et al. (2014), we only include sales following the first-predicted exposure, rather than from the start of the experiment, to increase the precision of our LATE estimates. Table 3 shows that the ads increase site visits by 62,756 or 17.2% and sales by \$109,693 or 10.8%. The corresponding t-statistics and p-values are 13.29 ($p < 10^{-15}$) for site visits and 3.45 ($p < 10^{-3}$) 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 3, we see the ads cause 19,061 incremental visitors (26.6%), 1,016 incremental transactors (12.1%), and 1,132 incremental transactions (12.0%). These results are even more significant than their intensive-margin counterparts with t-statistics of 40.38 ($p < 10^{-15}$) for visitors, 5.91 ($p < 10^{-8}$) for transactors, and 5.42 ($p < 10^{-7}$) for transactions.

Though our Predicted Ghost Ad implementation passes all the tests in Section 5.2, we also compare our Predicted Ghost Ad estimates to the ITT estimates for robustness. Table 3 confirms that the ITT and LATE estimates are close and that a Hausman test for each outcome does not reject the LATE estimator.[11] Together with the validation checks in

[11]The reader can see that for site visitors, the ITT estimate's 95% confidence interval does not contain the LATE estimate. For both site visitors and transactors, the ITT estimates should be compared to the ghost ad estimates for outcomes during the campaign rather than after the first impression. The visits,

Table 2, we are confident that our Predicted Ghost Ad implementation delivers valid ad effectiveness estimates.

Our results not only demonstrate for the first time that retargeting can be effective, but do so with greater statistical significance than most experiments in the ad effectiveness literature. The result is important because retargeting is fraught with an endogeneity problem that makes 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 to a valid holdout group, we can show this retargeting campaign creates a large increase in both site visits and sales. This baseline comparison with a control group differentiates our results from the retargeting studies by Lambrecht & Tucker (2013) and Bleier & Eisenbeiss (2015). These studies instead compare retargeting campaigns that feature more or less personalized information regarding the focal product. The studies only agree that content matters for personalization, which improves ad effectiveness when the user browses product-related content (Lambrecht & Tucker, 2013) or shopping sites (Bleier & Eisenbeiss, 2015). Finally, ours is the only study to examine the impact on sales revenues among these studies.

The Ghost Ad methodology changes the economics of acquiring ad effectiveness information for advertisers. In our study, a mid-size online retailer obtains highly significant results in spite of 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 Ads 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

transactions, and sales variables are additively separable before and after the first predicted impression, whereas the visitor and transaction indicator variables are not. Thus, the expected treatment effects for the number of unique visitors and transactors differ by the time period of the outcome. For a proper comparison, the visitor and transactor Hausman tests compare the ITT estimates to the predicted ghost ad estimates using during-campaign outcomes: the estimated treatment effects are 13,106 (s.e. 575) for visitors and 998 (s.e. 174) for transactors.

experiments that treat a small proportion of eligible users, 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 treated users. Sportsing can obtain the same statistical power as the original test by running a concentrated test that doubles the per user ad spend while shrinking the treatment group to a 6/94 design, if the ad effect is proportional to ad spend.[12] 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.

Hence, the Ghost Ads methodology improves upon the pessimistic outlook on measuring the returns to advertising in Lewis & Rao (2015). Lewis & Rao (2015) 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 is caused by 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 approach decreases the variance in the estimates relative to ITT. The ratio of these estimates' variances in Table 3 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.[13]

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

[12]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 $= \delta / SE(\delta)$ to be the same when the treatment effect doubles, we find that the proportion of users in the treatment group $p = 5.6\%$ solves $2\delta / \frac{\sigma}{\sqrt{Np(1-p)}} = \delta / \frac{\sigma}{\sqrt{N*0.3*0.7}}$. While we are assuming no ad wear-out (i.e., diminishing returns) from doubling the ad spend, this is consistent with the findings of Lewis (2010) and Johnson et al. (2014).

[13]The experiment also enjoyed better statistical power from a larger (30%) control group, concentrated average ad frequency (19 impressions per user), and directly-linkable online-only purchase channel. However, this retailer typifies those in Lewis & Rao (2015) with a similar coefficient of variation on sales of $\frac{\sigma}{\mu} = 10.1$ during this two-week campaign.

interim. Moreover, a cookie-switching consumer that re-enters the experiment could switch treatment groups, further attenuating the measured ad lift. The median cookie age in our data is about 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. Hence, if a user's ad exposures are linked to one cookie but 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 the absolute ad lift estimates as lower bounds.

6 Implementation challenges

The idea of Predicted Ghost Ads seems simple, but we have spent three years building the implementation at Google which now records more than 100 million predicted ghost ads each day, and the work is ongoing. In this section, we discuss some of the challenges, lessons learned, and possible extensions for the methodology in the online display ad setting.

The engineering challenge with the Predicted Ghost Ad methodology is that its implementation must be perfectly symmetric between treatment groups or else the experimental comparison could be biased. Perfection is elusive because ad platforms—at least in the online display advertising setting—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. In Section 4, we explain that we abandoned the pure Ghost Ad approach and embraced the Predicted Ghost Ad methodology to address the ad platform's inability to control ad delivery. Section 4 also raises the problem of externalities across multiple ghost ad campaigns. We use 'auction isolation' to eliminate these externalities and scale the Predicted Ghost Ad methodology.

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 either be won by a single display ad or by multiple separate AdWords text 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 et al., 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. Though 'auction isolation' avoids collisions between ghost ads here, the predicted ghost ad system needs to be implemented over all impressions on a page to work.

Some problems arise when an advertiser runs other campaigns that compete with its ghost ad experiment. When this happens, a concurrent campaign can drive up the cost per impression of the treatment campaign. Worse, the delivery of the concurrent campaign could also differ by treatment group. This can happen if the ad platform substitutes the 'withheld' treatment ads for the concurrent campaign's ads in the control group or if the ad platform learns from the treatment campaign's performance, which informs the concurrent campaign's delivery.[14] Thus, the experiment would not hold fixed the advertiser's other ads, and the PGA LATE estimates for the treatment campaign would need to account for the concurrent campaign's interference. This problem can also arise when the advertiser runs a concurrent campaign through an intermediary that purchases from the ad platform. In this case, the ad platform may by unable to ascertain the problem because the intermediary may not share its advertiser's identity. Cross-campaign interference is a fundamental problem that can be avoided by running one campaign at a time or targeting each campaign to distinct groups of users. For instance, the Predicted Ghost Ad approach can compare the effectiveness of two campaigns (a copy test) by creating three randomly assigned groups of users: campaign A

[14]Some users will be exposed anyways to the concurrent campaign, so they are in a sense what Angrist et al. (1996) call the 'always-takers' in LATE. The difference here is that the intensity of treatment in the experiment could vary so that that the always-takers in the control group will see fewer ads on average than their counterparts in the treatment group.

treatment, campaign B treatment, and a *single* control reused *twice* as a control group with ghost ads for *each* campaign.[15]

We can also modify the Predicted Ghost Ad methodology to record information on ad viewability. A viewable impression is an ad that appears in the user'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 wish to measure the impact of viewable ads since non-viewable ads logically have little or no effect. However, the ad platform only has viewability information for the ads it serves and not those served by another intermediary. Thus, while the ad platform has viewability information for the treatment ads, the viewability data is incomplete for ads shown to the control group. We 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 treatment ad. To operationalize this, we can run simulated auctions both with and without the treatment ad for all users and record the predicted ghost ad and predicted displaced ad. In this way, we can split the predicted displaced ads (conditional on predicted ghost ad exposure) by whether they record viewability and do so symmetrically across treatment groups.

7 Conclusion

Our Ghost Ad and Predicted Ghost Ad methodologies promise to revolutionize ad effectiveness measurement by changing the economics of experimentation. In eliminating the costs of PSAs, ghost ads make experimentation accessible to advertisers at the mere press of a button. This should encourage more advertisers not only to start experimenting and but

[15] An ad platform's endogenous ad serving can bias the direct comparison of different ad copy in the same way as the comparison with a PSA. Pre-experiment randomization checks of the outcomes can help detect such bias. Copy tests can mitigate this problems for instance by using CPM pricing, but to an unknown extent without a control group. With predicted ghost ads, we can reuse the same control group for alternative ad copy groups which increases power over splitting the sample into separate control groups per copy or using Intent-to-Treat. This also improves generalizability of a copy test's results by allowing the ad system to optimize the performance of each creative by delivering each to its respective most responsive audience.

also to make experimentation routine. 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 or even ad platforms will use *explore-exploit tests*, in which advertisers explore the effects of different campaigns on small groups of users before exploiting the best campaigns on the rest. Hundreds of advertisers have used Google's implementation of the system, which delivers millions of experimental ad impressions daily.

Ghost ads reduce the cost of running experiments, and wide-spread experimentation enables not only marketers to better allocate their ad budgets but also academics to refine their ad models. In particular, meta-studies of ghost ad experiments will shine new light on ad attribution and ad stock models. In the future, we expect the Ghost Ad methodology to evolve further. For instance, ghost ads can be combined with ad viewability technology to better measure the effect of a viewed impression by ignoring ads that appear only for a split-second or remain off-screen 'below the fold.'

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, the Ghost Ad methodology can be applied to other forms of advertising beyond online display ads. Search ads are a natural candidate since they are sold using similar digital infrastructure. Other media could follow since programmatic ad buying and ad measurement is spreading to media like television and radio. In the future, we predict new advertising media experiments will employ the Ghost Ad methodology to improve the theory and practice of marketing.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455.

Authors (2015). The online display ad effectiveness funnel & carry-over: A meta-study of ghost ad experiments.

Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World wide web* (pp. 283–292).: ACM.

Bakshy, E., Eckles, D., Yan, R., & Rosenn, I. (2012). Social influence in social advertising: evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (pp. 146–161).: ACM.

Bart, Y., Stephen, A. T., & Sarvary, M. (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*.

Blake, T., Nosko, C., & Tadelis, S. (2013). Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *NBER Working Paper*, (pp. 1–26).

Bleier, A. & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, Forthcoming.

Candela, J. Q., Bailey, M., & Dominowska, E. (2014). Machine learning and the facebook ads auction. In *A joint session for EC, NBER and Decentralization on CS and Economics*: EC '14.

Evans, D. (2009). The online advertising industry: Economics, evolution, and privacy. *Journal of Economic Perspectives*, 23(3), 37–60.

Goldfarb, A. & Tucker, C. (2011a). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3), 389–404.

Goldfarb, A. & Tucker, C. (2011b). Privacy regulation and online advertising. *Management Science*, 57(1), 57–71.

Google (2015). Where ads might appear in the display network.

Hoban, P. R. & Bucklin, R. E. (2015). Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research*, 52(3), 375–393.

Hu, Y., Lodish, L. M., & Krieger, A. M. (2007). An analysis of real world TV advertising tests: A 15-year update. *Journal of Advertising Research*, 47(3), 341.

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

Johnson, G. (2013). The impact of privacy policy on the auction market for online display advertising. *Available at SSRN 2333193*.

Johnson, G., Lewis, R. A., & Reiley, D. H. (2014). Location, location, location: Repetition and proximity increase advertising effectiveness.

Kalyanam, K., McAteer, J., Hodges, J., & Lin, L. (2015). Cross channel effects of search engine advertising on brick and mortar retail sales: Insights from multiple large scale field experiments on google.com.

Lambrecht, A. & Tucker, C. (2013). When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research*, 50(5), 561–576.

Lavrakas, P. J. (2010). *An evaluation of methods used to assess the effectiveness of advertising on the internet*. Technical report, Interactive Advertising Bureau.

Lewis, R. & Nguyen, D. (2015). Display advertising's competitive spillovers to consumer search. *Quantitative Marketing and Economics*, 13(2), 93–115.

Lewis, R., Rao, J. M., & Reiley, D. H. (2015). Measuring the effects of advertising: The digital frontier. In A. Goldfarb, S. M. Greenstein, & C. E. Tucker (Eds.), *Economic Analysis of the Digital Economy*. University of Chicago Press.

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

Lewis, R. A. (2010). *Where's the "Wear-Out?": Online Display Ads and the Impact of Frequency*. PhD thesis, MIT Dept of Economics.

Lewis, R. A. & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics (forthcoming)*.

Lodish, L., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., & Stevens, M. (1995). How TV advertising works: A meta-analysis of 389 real world split cable TV advertising experiments. *Journal of Marketing Research*, 32(2), 125–139.

Sahni, N. (2015). Advertising spillovers: Field experimental evidence and implications for returns from advertising.

Sahni, N. & Nair, H. (2015). Native advertising: Evidence from mobile advertising field experiment. Presented at Marketing Science (Baltimore).

Simester, D., Hu, J., Brynjolfsson, E., & Anderson, E. (2009). Dynamics of retail advertising: Evidence from a field experiment. *Economic Inquiry*, 47(3), 482–499.

Yildiz, T. & Narayanan., S. (2013). Star digital: Assessing the effectiveness of display advertising. *Harvard Business Review: Case Study*.

Figures & Tables

Figure 1: The Ideal Experimental Design

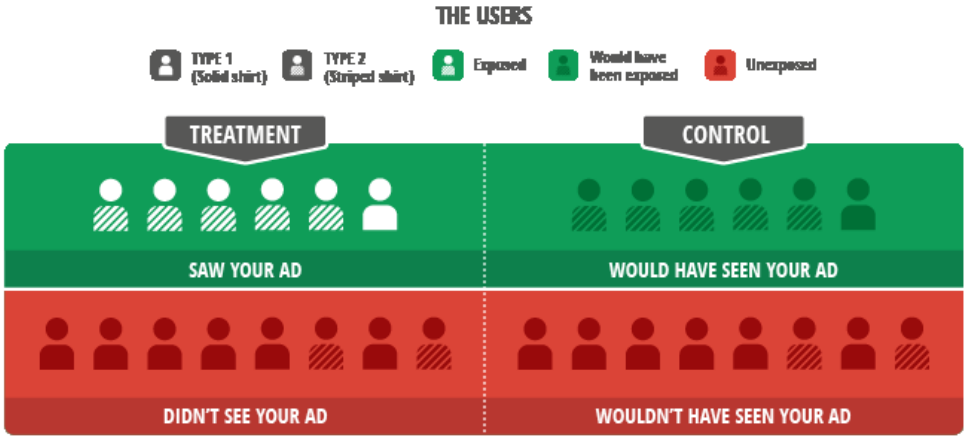

Figure 2: Intent-to-Treat Experimental Design

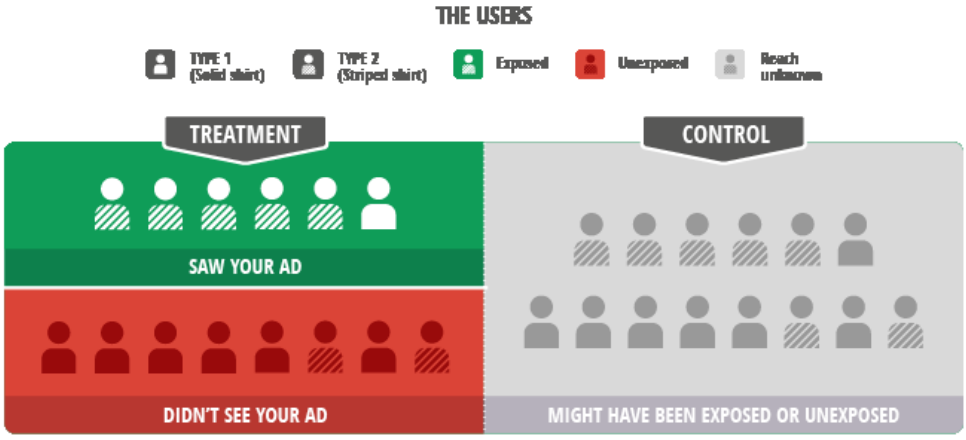

Figure 3: Biased PSA Experimental Design

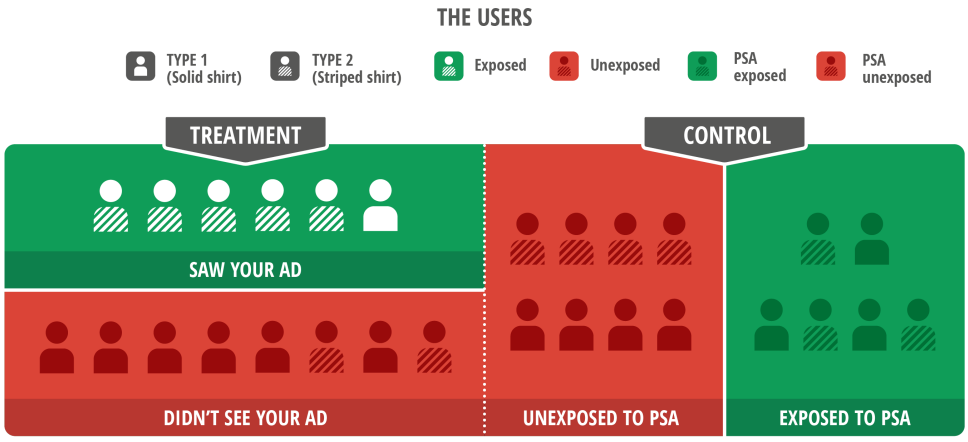

Figure 4: Ghost Ads in Action

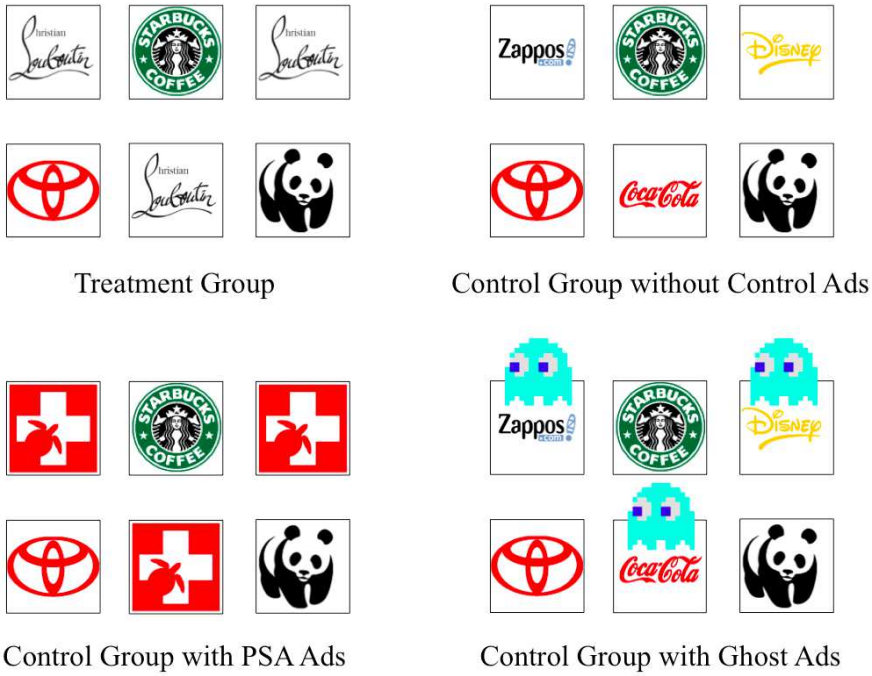

Figure 5: Ghost Ads in Three Generations of Ad Platforms

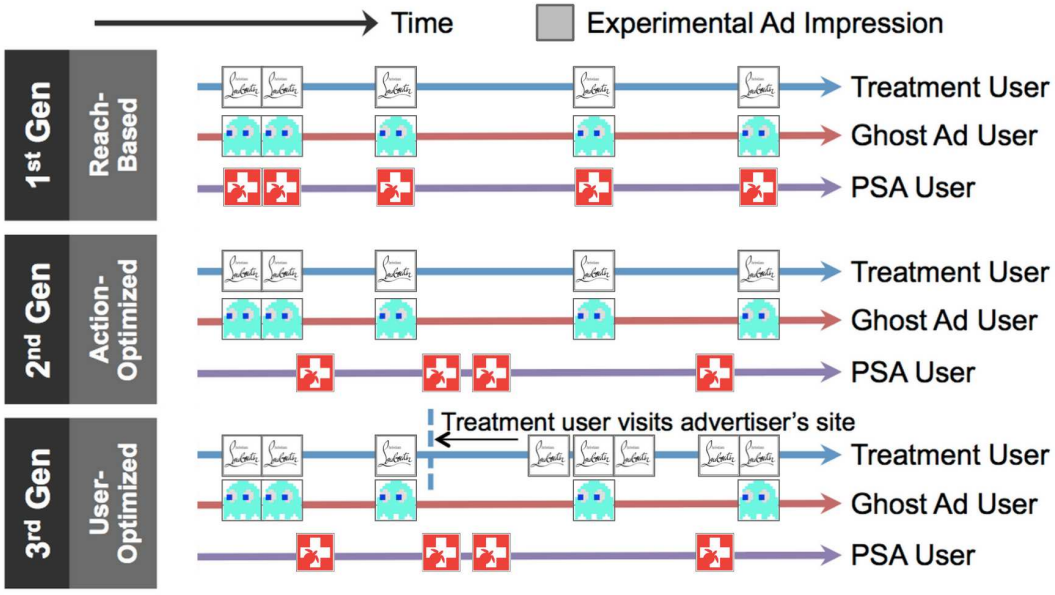

Figure 6: Predicted Ghost Ad Flow Diagram

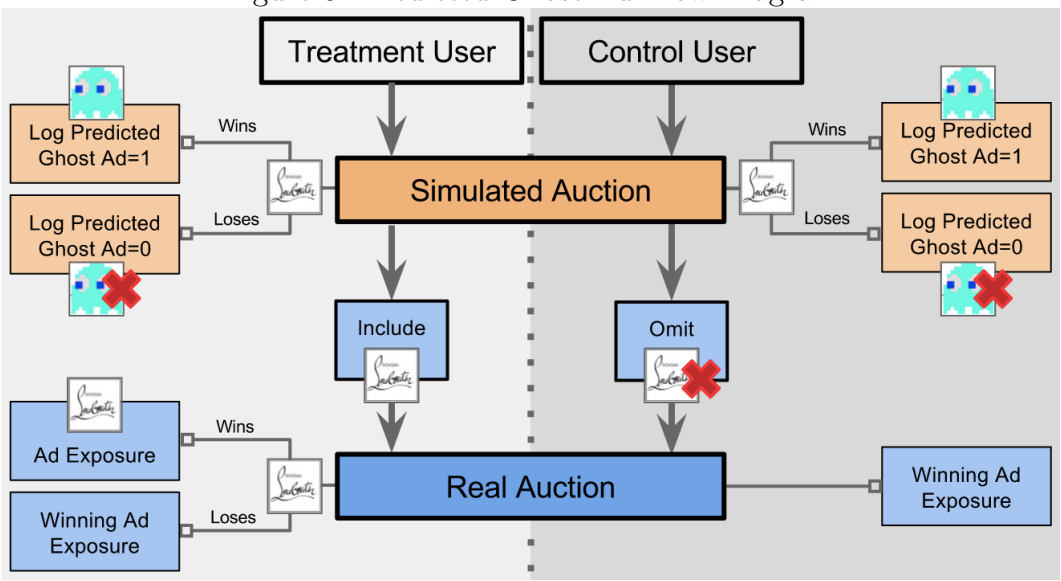

Figure 7: Retargeting Creative Resembling Those in Experiment

Table 1: Comparison of Experimental Approaches as Ad Platforms Evolve

			Does experimental approach work?	
Generation	Ad Platform	Campaign Capability	PSA	Ghost Ad
1	Reach-based	CPM	√	√
2	Action-optimized	CPC, CPA		√
3	User-optimized	Retargeting		√ (First Ad)

Table 2: Experimental Validation for Predicted Ghost Ad Exposed Users

	Treatment/Control Test of Equality *p-values throughout*		Test
Demographics	Treated Users		
Country	0.85		Chi^2
City	0.50		Chi^2
Predicted Ghost Ads	First Impression*	Average Impression†	
Creative Shown	0.68	-	Chi^2
Publisher Website	0.49	-	Chi^2
Predicted Click-Through Rate	0.47	0.00	t-test
Predicted Conversion Rate	0.10	0.00	t-test
Predicted Cost per Impression	0.10	0.00	t-test
Cost per Click	0.38	0.00	t-test
Pre-Treatment Outcomes	30 days prior to experiment	During experiment prior to exposure	
Site Visits	0.92	0.33	t-test
Transactions	0.76	0.21	t-test
Sales	0.75	0.11	t-test

Notes: Tests are between 396,793 predicted exposed users in the treatment group and 169,584 predicted exposed users in the control group. *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. †Not expected to be equivalent between treatment and control. Averages taken across impressions at the user level.

Table 3: Ad Effectiveness Results for Dynamic Remarketing Campaign

	(1) Site Visitors	(2) Site Visits	(3) Transactors	(4) Transactions	(5) Sales
Predicted Ghost Ad LATE Estimates					
Post-first predicted impression outcomes					
Treatment	90,841	426,984	9,409	10,606	$1,124,140
Control*	71,780	364,228	8,393	9,474	$1,014,447
Difference	19,061	62,756	1,016	1,132	$109,693
	(472)	(4,722)	(172)	(209)	(31,799)
t-statistic	40.38	13.29	5.91	5.42	3.45
% Lift	26.6%	17.2%	12.1%	12.0%	10.8%
Difference/Exposed	0.048	0.158	0.0026	0.0029	$0.28
Intent-to-Treat Estimates					
During campaign 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)	0.611†	0.584	0.468†	0.209	0.223
Rel. Variance of ITT	14.5	8.5	5.9	8.7	17.1

Notes: Results employ experimental difference estimator. Standard errors are in parentheses; ITT standard errors use a conservative Poisson estimator for $Var(\Sigma x) \approx \Sigma x^2$ as the population size is unknown due to the experiment's 'on the fly' randomization. *Control group total outcomes rescaled by 7/3 to match treatment groups in 70%/30% treatment split. †Hausman test compares consistent ITT estimates with efficient Predicted Ghost Ad LATE using during-campaign outcomes for a proper comparison (see footnote 11).