The Online Display Ad Effectiveness Funnel & Carryover: A Meta-study of Predicted Ghost Ad Experiments

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

September 23, 2016

Abstract

We present a meta-study of 432 online display ad field experiments on the Google Display Network. The experiments feature 431 advertisers from varied industries and on average include 4 million users and last 20 days. These experiments employ the Predicted Ghost Ad experimentation methodology. Across experiments, we find median lifts of 16% in site visits and 8% in conversions. We relate the baseline attrition as consumers move down the purchase process—the marketing funnel—to the incremental effect of ads on the consumer purchase process—the 'ad effectiveness funnel.' We find that incremental site visitors are less likely to convert than baseline visitors: a 10% lift in site visitors translates into a 5-7% lift in converters. We then examine the carryover effect of the campaigns to determine whether the in-campaign lift carries forward after the campaign or instead only causes users to take an action earlier than they otherwise would have. We find that most campaigns have a modest, positive carryover four weeks after the campaign ended with a further 6% lift in visitors and 16% lift in visits on average, relative to the in-campaign lift.

Keywords: Field experiments, advertising effectiveness, meta-study, digital advertising

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

1 Introduction

Since their inception in 1994, online display ads have grown to a $32 billion industry in the United States including mobile and video ads (eMarketer, 2016). Despite this, we know few general facts about the medium's effectiveness that have been established by meta-studies of field experiments or cross-advertiser studies. For instance, we know that online display ads increase brand metrics in surveys (Goldfarb & Tucker, 2011; Bart et al., 2014), that ad frequency effects are heterogeneous (Lewis, 2010; Sahni, 2016), and that ads create spillovers for competitor conversions and searches (Lewis & Nguyen, 2015; Sahni, 2016). We provide novel evidence from 432 experiments that online display ads increase user site visits and conversions for the advertiser—by a median of 16% and 8% respectively. We find that incremental site visitors from the campaign are half as likely to convert as the baseline site visitors. Finally, we find modest carryover in ad effectiveness four weeks after the campaign.

In this paper, we present a meta-study of 432 large-scale field experiments on the Google Display Network (GDN). On average, these experiments reach over 4 million users and last 20 days. The experiments include hundreds of advertisers from a variety of industries. GDN is one of the largest online display ad platforms encompassing 2 million websites and reaching over 90% of global users (Google, 2015). Though GDN implements the experiments, the advertisers devise all aspects of the campaign including: ad creatives, campaign tactics, target audience, spend, duration, and outcome measures. GDN's experimentation platform uses the Predicted Ghost Ad methodology developed by Johnson et al. (2016a). Predicted Ghost Ads deliver valid ad effectiveness estimates even when the advertisers use computer algorithms to optimize in-campaign performance. As GDN makes its experimentation platform available to advertisers for free, we were able to collect hundreds of studies in under four months. To protect the privacy of advertisers, Google did not share with us any information

about the advertisers, campaigns, creatives, or exposed users beyond the experimental lift.

We examine the impact of online display ads on online outcomes that affect the advertiser's bottom line. We measure the visits to the focal advertiser's website by users in the experiment. We also measure the users' conversions, which the advertisers define and can include: purchases, downloads, sign-ups, and key page-views. Relative to the baseline established by the control group, site visits are 16% higher and conversions are 8% higher on median. Across studies, the estimated lift in site visits is positive and significant at the one-sided $\alpha = 0.025$ level in 56% of cases and 29% of cases for conversions. In this sense, the studies we consider have greater statistical power than the split-cable TV meta-study in Lodish et al. (1995a) where the majority of sales lift estimates were statistically insignificant at the one-sided $\alpha = 0.20$ level. Our studies collectively establish with high statistical significance that online display ads cause users to alter their online behavior: binomial tests reveal increases in both site visits ($p = 7.4 \times 10^{-213}$) and conversions ($p = 2.9 \times 10^{-40}$).

The marketing funnel refers to the attrition in the consumer purchase process as consumers move through stages like awareness, interest, desire, and then action. Academics and practitioners often observe upper funnel outcomes—like survey measures of brand favorability, searches, and site visits—rather than the lower funnel outcomes that most interest marketers—like sales. We estimate the *elasticity of the ad effectiveness funnel*, which we define as the ratio between the percentage lift at higher and lower stages of the marketing funnel. The elasticity of the ad effectiveness funnel also tells us whether incremental users are more or less likely to progress down the marketing funnel than baseline users. We find that the incremental users exhibit greater attrition: a 10% lift in site visits translates into about a 5-7% lift in conversions. Marketing practitioners can use this 'rule of thumb' as a point of departure to better optimize their campaigns especially because the effect of online display ads on purchases is imprecisely estimated (Lewis & Rao, 2015) if purchases are observed at all.

We also use our meta-study to quantify the long-run or carryover effect of ad campaigns.

Ad models often assume that the carryover effect of advertising is positive and decays geometrically over time (see e.g. Nerlove & Arrow 1962). However, some empirical evidence suggests that carryover can be negative; that is, ads merely cause consumers to substitute their actions forward in time (Simester et al., 2009). Advertising practitioners must know whether short-term ad effectiveness estimates understate or overstate the long-term effects of the ads to allocate their ad budgets. We find that, four weeks after a campaign, incremental visits rise a further 16% on average beyond the during-campaign lift and incremental visitors increase a further 6%. Though the median carryover effect is positive, our carryover estimates suggest heterogeneity in the sign of the ad campaign carryover effects.

The next section describes our methodology for measuring ad effectiveness in field experiments. Section 3 describes our sample of ad studies. Section 4 presents the distribution of ad effectiveness estimates, the ad effectiveness funnel elasticity estimates, and the carry-over estimates. Section 5 concludes.

1.1 Literature review

Meta-studies of advertising field experiments reveal empirical regularities that case studies cannot. Meta-studies remain rare due to the high cost of ad experiments especially in offline media. Lodish et al. (1995a) stand out for examining 389 'split-cable' television ad experiments. These experiments are either weight tests that compare different ad intensities or copy tests that compare different ad creatives. The majority of these experiments are not statistically significant (p=0.2, one-sided) likely because they only involve thousands of viewers and measure the effect of ads on sales, which contribute to low statistical power.

Digital advertising has facilitated large-scale experiments with user-level randomization of advertising exposure. Though purchase data remain rare, these experiments often measure survey outcomes or online outcomes like site visits and conversions. In the past, field experiments have been uncommon even in online display advertising due to the high costs of setting up experiments and purchasing control ads (Gluck, 2011). Recently, several major on-

line display ad platforms—Facebook, Twitter, and Google—introduced free experimentation platforms for advertisers, which induced thousands of advertisers to run field experiments (Shrivastava, 2015; Facebook, 2016).

Focusing on online display ads, meta-studies of field experiments have begun to uncover some facts about the medium's effectiveness. Goldfarb & Tucker (2011) study almost 3,000 tests averaging 900 survey-takers each and find that targeted and obtrusive ads increase purchase intent, but the two combined do not. Bart et al. (2014) examine 54 studies with 740 users on average and show that mobile display advertising is effective for low involvement products. Lewis (2010) documents heterogeneity in ad frequency effects on clicks using a natural experiment with 30 studies and an average of 40 million users. Lewis & Nguyen (2015) use the same natural experiment to examine the competitive effects of banner advertising on online consumer search. Gordon et al. (2016) examine a dozen experiments averaging 36 million eligible users on Facebook and show that several observational methods used in industry are unreliable in recovering the experiment's causal estimate. Lewis & Rao (2015) examine 25 experiments at Yahoo! that reached over a million users on median to quantify the statistical power problem in this setting: the effect of ads are small in relation into the noise in the user outcome data. Large cross-advertiser experiments accomplish the same goal of generalizability as do meta-studies. Bakshy et al. (2012) show that social cues improve clicks and likes on Facebook ads. Sahni (2015, 2016) examine a hybrid search advertising setting on a website where consumers search for restaurants, but the search ads are banners rather than text ads. Sahni (2015) examines the effect of temporal spacing between exposures while Sahni (2016) measures the competitive effects of the ads.

To our knowledge, we are the first to measure the elasticity of the ad effectiveness funnel in online display advertising across a large number of field experiments. To do so, we compare the incremental upper funnel lift in site visits to the lower funnel lift in conversions. Cross-advertiser studies in restaurant search platforms provide conflicting evidence: Sahni (2015) shows that sales leads (48%) increase by more than site visits (15%), whereas Dai &

Luca (2016) show that lower funnel outcome lifts are less than the upper funnel lift. For online display ads, Johnson et al. (2016a) find in one field experiment with retargeting that transactions (12%) increase by less than site visits (17%). In a related online display ad field experiment, Hoban & Bucklin (2015) instead document heterogeneous ad effects on site visits by the user's stage in the purchase funnel prior to the experiment.

The marketing literature has long wrestled with the long-run effects of advertising. As Sethuraman et al. (2011) indicate, most of the time-series empirical ad effectiveness literature assumes an ad stock model to link sales to current and past advertising. In these ad stock models, the impact of past advertising decays geometrically at a constant rate (see e.g. Nerlove & Arrow, 1962; Clarke, 1976). While the literature often uses ad stock models to measure the long-run effect of ads, some evidence suggests that ads might have a *negative* effect in the long run. In particular, ads may cause consumers to purchase earlier than they otherwise would have, so that consumers substitute their purchases forward in time. Simester et al. (2009) document this phenomenon empirically in their catalog-mailing experiment. On the other hand, Sahni (2015) detects a positive and significant carryover on the users' subsequent browsing session, but finds that a memory-based model better captures the data than an ad stock model. Lodish et al. (1995b) examine 55 television field experiments and show that the long-run effects of ads on sales are positive on average two years after the tests. We return to this important question with hundreds of tests averaging millions of users each.

2 Methodology

In Section 2.1, we lay out the logic of an ad effectiveness experiment. We describe our default Predicted Ghost Ad methodology in Section 2.2 and the meta-study's fallback Intent-to-Treat method in Section 2.3. For a detailed discussion of these methodologies and the assumptions that underpin them, see Johnson et al. (2016a).

2.1 Experimental ad effectiveness measurement

Online marketers wish to measure the effect of their advertising campaigns on user outcomes. Online marketers wish to know: how do the users who are exposed to my campaign react compared to if I had not advertised? A holdout experiment answers this question by randomly selecting users for a control group who are held out from exposure to the focal campaign.

In an experiment, users can be classified into four types by their treatment assignment and potential exposure to the focal campaign. Formally, user i is assigned treatment—denoted by Z_i—to either treatment ($Z_i = T$) or control ($Z_i = C$). Potential exposure D_i classifies user i by whether i *would be* exposed ($D_i = 1$) or *would not be* exposed ($D_i = 0$) to the focal ad campaign if i were assigned to the treatment group. The matrix below summarizes the four user types:

	Treatment	Control
Would be exposed	$Z = T, D = 1$	$Z = C, D = 1$
Would not be exposed	$Z = T, D = 0$	$Z = C, D = 0$

Viewed this way, the marketer wishes to compare outcomes among exposed users ($Z = T, D = 1$) to those among counterfactual exposed users ($Z = C, D = 1$), which correspond to the top-left and top-right groups in the above matrix. In other words, the marketer must compute the Average Treatment Effect on the Treated (ATET) for outcome y which is given by:

$$ATET = E\left[y|Z = T, D = 1\right] - E\left[y|Z = C, D = 1\right]. \tag{1}$$

Whereas the exposed users ($Z = T, D = 1$) are readily identifiable as those users in the treatment group who see one of the focal ads, the counterfactual exposed users ($Z = C, D = 1$) cannot be distinguished in this way. Johnson et al. (2016a) discuss three solutions: control ads, Intent-to-Treat, and the Ghost Ad and related methodologies. Below, we describe the

Predicted Ghost Ad and Intent-to-Treat approaches implemented at GDN.

2.2 Predicted Ghost Ads

Our meta-study's experiments apply the Predicted Ghosts Ad (PGA) methodology introduced by Johnson et al. (2016a) and implemented by GDN. The basic idea behind PGA is to approximate potential exposure D by predicting both the exposed and counterfactual exposed users. PGA's predicted exposed users are denoted by $\widehat{D}$ which approximates D. If $\widehat{D}$ is statistically independent from the treatment assignment Z, then $\widehat{D}$ enables a symmetric comparison between the treatment and control groups. In particular, the experimental difference among predicted exposed users is then a 'locally' valid ad effectiveness estimator for those predicted exposed users. To the extent that $\widehat{D}$ predicts the exposed users D, the PGA estimator will closely approximate ATET.

To construct $\widehat{D}$, Johnson et al. (2016a) suggest simulating the ad platform's ad allocation mechanism. Online display ad platforms usually employ a complicated auction to select an ad among many competing ads. Both treatment and control users enter a simulated ad auction that selects an ad among a set of eligible ads that include the focal ad. *Predicted ghost ad impressions* are those instances when the simulated auction selects the focal ad, which the ad platform records in a database. The predicted ghost ad impressions in this database define the binary variable $\widehat{D}$ which approximates D. The treatment and control users then enter the real auction to select which ad the ad platform will deliver, where the real auction only includes the focal ad in the set of eligible ads for treatment users. The outcome of the simulated auction has no bearing on the real auction. By construction, PGA is therefore independent from treatment assignment.

The Predicted Ghost Ad estimator is a Local Average Treatment Effect (LATE, see

Imbens & Angrist 1994) estimator given by:

$$LATE_{PGA} = \frac{E\left[y|\widehat{D}=1, Z=T\right] - E\left[y|\widehat{D}=1, Z=C\right]}{\Pr\left[D=1|\widehat{D}=1, Z=T\right]}. \tag{2}$$

In words, the numerator is the experimental difference between those treatment and control group users who are predicted to be exposed. The denominator scales up the experimental difference by the inverse conditional probability that treatment users are exposed given that they are predicted to be exposed. This probability is 0.999 in Johnson et al. (2016a) whose application also uses the GDN's PGA platform.

The LATE_{PGA} estimator is 'local' in the sense that it excludes users who are exposed to ads but are not predicted to be exposed. Thus, we can relate ATET and LATE_{PGA} as follows

$$ATET \quad = \quad LATE_{PGA} \cdot \Pr\left[\widehat{D}=1|D=1, Z=T\right] + \varepsilon \tag{3}$$

where ε captures the treatment effect arising from users who are not predicted to be exposed. Provided that $\widehat{D}=1$ captures almost all cases where $D=1$, ε is small. In Johnson et al. (2016a), GDN's predicted exposure excludes only 3.2% of exposed users and 0.2% of the campaign's ad impressions, so LATE_{PGA} approximates ATET well.

2.3 Intent-to-Treat

Intent-to-Treat (ITT) serves as the backstop methodology in our meta-study. The ITT approach provides valid ad effectiveness estimates as long as treatment assignment is random. ITT compares all eligible users in the treatment and control groups regardless of exposure. The ITT estimator is given by

$$ITT = E\left[y|Z=T\right] - E\left[y|Z=C\right]. \tag{4}$$

Imbens & Angrist 1994 tells us that ITT and ATET are related in expectation by

$$ATET_{ITT} = \frac{ITT}{\Pr\left[D = 1 | Z = T\right]}. \tag{5}$$

The intuition here is that the causal effect of the campaign arises only among exposed users (ATET), so that ITT will be the same in expectation after we rescale it by the inverse probability of exposure. However, the ATET_{ITT} estimator is less precise than the direct ATET estimator (see Lemma 2 in Johnson et al. 2016a). Now, by combining equations (5) and (3), we have

$$\frac{ITT}{\Pr\left[D = 1 | Z = T\right]} \quad = \quad LATE_{PGA} \cdot \Pr\left[\widehat{D} = 1 | D = 1, Z = T\right] + \varepsilon. \tag{6}$$

In Section 3, we use equation (6) to validate GDN's implementation of PGA. In Appendix A, we detail how we derive our ITT estimates and standard errors from the advertiser website's total activity.

3 Data

In this section, we describe our sample of experiments, the outcome variables, and a validation test of the Predicted Ghost Ad estimates.

3.1 Sample of Experiments

The experiments in this meta-study come from advertisers opting to use the Google Display Network's (GDN) experimentation platform. This experimentation platform has been in an 'alpha' development stage since it was launched in 2014. Google does not advertise the platform on the GDN webpage nor does Google make this platform available to all advertisers. To use the experimentation platform, an advertiser must interact with GDN's

salesforce. The only criterion for participating is that the advertiser must have a medium- or large-sized budget. In view of the power problems described in Lewis & Rao (2015), we felt that small advertisers would find the experimental results too imprecise to be helpful. The advertisers in our sample are a diverse and self-selected group that is interested in ad measurement and experimentation. Unfortunately, we were unable to obtain data from Google on the advertisers, the campaigns or the exposed users. Thus, we are unable to correlate ad effectiveness with advertiser, campaign, or user attributes.

The GDN experiments randomized treatment assignment at the user level. A user is identified as a unique cookie on a browser-device combination. GDN uses this cookie to track a user's exposures to ads across publishers and—with the aid of advertisers—track the user's subsequent interactions with the advertiser's website. A consumer may have multiple cookies corresponding to multiple devices, which attenuates ad estimates if users take incremental actions on another device than the device where they see the ad. Also, some users will be assigned to multiple treatment groups if the same campaign could reach them on different devices, which also attenuates ad effectiveness estimates (Coey & Bailey, 2016). Nonetheless, marketers must use cookies as the unit of analysis if the ad platform does not track logged-in users (see e.g. Bleier & Eisenbeiss 2015; Hoban & Bucklin 2015; Lambrecht & Tucker 2013).

Our sample only includes experiments where we expect 100 or more users in the treatment group to trigger an outcome under the null hypothesis that the ads are ineffective.[1] By restricting our sample in this way, we avoid tests with few users, advertisers with very low baseline activity, narrowly defined outcome variables, and potential outliers in the relative lift estimates. Our meta-study includes 432 experiments, out of the 606 experiments in the sample collection period. On occasion, we will refer to the subsample of *powerful studies*, which we define as those experiments which exceed our statistical power threshold. We set this threshold such that a 5% (one-sided) test should reject a zero ad effect 95% of the time

[1]The control group establishes the no ad effect baseline. Given the 70/30 split between treatment and control, 100 users triggering an outcome corresponds to 43 (100*3/7) actual occurrences in the control group.

when the alternative hypothesis is a 100% increase over the control group baseline. This threshold means that the control group's outcome must be at least 3.3 times larger than the standard error of the experimental difference estimate.

Our study pulls experiments after a major update of GDN's experimentation platform that improved performance. Johnson et al. (2016a) describe the 'auction isolation' feature that underlies this update. Our sample collection begins on June 10, 2015 and ends on September 21, 2015. 28% of the experiments were still in progress on September 21 and are cut short. The experiments in our sample last an average of 20 days and range from 1 to 102 days long. Experiments include an average of 4 million predicted-exposed users (in the treatment and control groups) with tests ranging from 21 thousand to 181 million users. The experiments assign 70% of users to the treatment group and 30% to the control group. In this three month sample, only one advertiser had two experiments and the remaining experiments are unique to individual advertisers.

3.2 Outcomes

Our outcome variables are site visits and conversions as defined and set-up by the advertisers themselves. These outcomes are recorded using tiny images called pixels that allow the advertiser and the platform to know which users visit which parts of the advertiser's website. To do this, the advertiser places pixels on some or all of the advertiser's webpages to capture user visits and designates 'conversion' pixels that record key pageviews. Conversions may include purchases, sign-ups, or store location lookups. In half of the studies, the advertiser tracks multiple outcome variables using groups of pixels, but these pixel outcomes are not clearly labeled on the ad platform side. We choose a single site visit and conversion outcome for each test according to some rules. We select the outcome with the greatest number of baseline users who trigger the outcome in order to find the most broadly applied site visit or conversion pixel. We break ties with the largest number of triggered pixel outcomes in the baseline. By selecting a single pixel outcome, our absolute ad effectiveness estimates will be

conservative both because they might exclude some site visits and conversions and because the campaign's goal may not correspond to the selected outcomes. Also, selecting a single pixel outcome avoids the risk of double-counting outcomes if we were to instead sum them up. We drop the site visit pixel outcome whenever it duplicates a conversion pixel. Note that some studies have either a site visit or a conversion outcome, but not both.

Following Johnson et al. (2016b), we refine the outcome variables in the PGA estimates by omitting realized outcomes that occur prior to the first predicted exposure. The logic here is that the campaign cannot affect a user before the user has been exposed to the first ad. Johnson et al. (2016b) show that post-exposure filtering improves the precision of the standard errors of their experimental estimates by 8%.

3.3 Validation of PGA Implementation

We test the performance of GDN's implementation of PGA by comparing our LATE_{PGA} estimates to our unbiased ITT estimates using equation (6) for site visits and conversions. We use a Hausman test which evaluates the consistency of the PGA_{LATE} estimator compared to the consistent, but less efficient ITT estimator. The Hausman test could reject if the predicted ghost ads ($\hat{D}$) are not independent of treatment assignment Z or if under-prediction (ε in eq. 6) is too large. In 95% of the experiments, the Hausman test does not reject the null hypothesis that the LATE_{PGA} and ITT estimates are identical at the 5% level.[2] The 5% failure rate here is consistent with false positives. Nonetheless, we fall back on ITT estimates in our analysis whenever the Hausman test rejects.

[2]Since a small fraction of eligible users are typically exposed, ITT estimates can have as much as 10 times higher variance than the PGA-LATE estimates in this setting (Johnson et al., 2016a), which makes for a weaker test.

4 Meta-Analysis Results

In the three subsections below, we respectively examine the overall results across all tests, the elasticity of the ad effectiveness funnel, and the carryover effect of advertising.

4.1 Overall Treatment Effects

We begin by presenting the overall ad effectiveness estimates across our 432 tests. Of these, we have site visit data for 347 tests and conversion data for 184 tests. We also distinguish between test that meet the minimum statistical power threshold described in Section 3; the *powerful* studies include 339 site visit tests and 174 conversion tests.

To compare disparate tests, we normalize a test's lift by the baseline outcome, which is established by the control group. As we saw in Section 3, the 432 tests vary in many dimensions including the campaign's duration and reach. To make these comparable, we normalize by the baseline activity which should also be proportional to the campaign's duration and (counterfactual) reach. For our default PGA estimator, the relative lift is given by

$$y_{\Delta} \equiv \frac{E\left[y|Z=T,\hat{D}=1\right]-E\left[y|Z=C,\hat{D}=1\right]}{E\left[y|Z=C,\hat{D}=1\right]}.$$

The relative ITT lift estimator is similar (see Appendix A). This normalization means that the relative lift can be very large when the denominator is small, for instance for new advertisers or narrowly defined outcome measures (Section 3.2). For this reason, we restrict attention to studies where we expect the baseline to register at least 100 occurrences of the outcome variable (Section 3.1).

Table 1 summarizes the lift estimates across all experiments and across the more statistically powerful experiments. Across all experiments, we see a median increase of 16.6% in site visits with a 10%-90% interquantile range of [-1.1%, 215.6%]. The simple average and average weighted by number of predicted exposed users are both high at about 1200 and

800% due to influential outliers. The lift in conversions is smaller with a median increase of 8.1% and a 10%-90% interquantile range of [-8.9%, 83.4%]. The average and weighted average conversions are a more modest 19.9% and 25.3%. On the extensive margin, we see a median of 21.8% incremental visitors and 7.9% incremental converters. Turning to the powerful studies, we see that the average and weighted average of incremental site visits plummet to 75% and 113%, which indicates that outliers among the low-powered studies were driving the average.

Figures 1 and 2 illustrate the variability of the incremental visit and conversion point estimates across tests as well as the variability of the estimates themselves. The 95% confidence whiskers are wide in tests with small samples or tests that employ the ITT rather than LATE_{PGA} estimator. As the histograms in the sidebars illustrate, the majority of the point estimates are positive for both visits (85%) and conversions (74%). The estimates are so variable that they give rise to negative point estimates and even cases that are negative and individually significant. Fewer that 2.5% of tests are negative and significant at the 2.5%, one-sided level, which is consistent with false positives. As Table 1 shows, the overall weight of evidence suggests a positive and jointly significant effect of advertising. Of 347 site-visit tests, 202 are significantly different from zero at the 5% two-sided level, and only 7 of those point estimates are negative. Of the 184 conversion tests, 57 are significant, and only 4 point estimates are negative.

This meta-study represents strong collective evidence that ads affect real-world consumer behavior. We use a binomial test to evaluate the joint statistical significance of the findings. Table 1 shows that 195 of the 347 site visit lift estimates are significant at the $\alpha = 0.025$ one-sided level. The binomial test compares these 195 successes out of 347 trials to the null hypothesis of a 2.5% success rate for meeting that p-value threshold due to random chance. For visits, the binomial test returns a p-value of $7.4 * 10^{-213}$. Table 1 shows that across the other outcomes, the p-values vary between $p = 6.5 * 10^{-289}$ and $p = 1.8 * 10^{-34}$. Taken together, these studies establish that online display ads increase both visits and conversions.

This collective evidence compares favorably to the split-cable TV meta-studies. In Lodish et al. (1995a), 39% of the 292 weight tests have $p \leq 0.2$, one-sided. In a follow-up meta-study by Hu et al. (2007), 30% of the 210 weight and hold-out tests have $p \leq 0.015$, one-sided. We compute analogous binomial test p-values of $p = 2.3 * 10^{-14}$ for the weight test in Lodish et al. (1995a) and $p = 4.3 * 10^{-62}$ for the follow-up study by Hu et al. (2007).[3] To be clear, these comparisons do not mean that online display ads 'work' better than TV ads, rather we suspect our lift estimates are more precisely estimated because our sample sizes are much larger. Whereas our average sample size is 4 million, the split-cable TVs studies have on the order of 10,000 households. Nevertheless, the present meta-study differs from the split-cable meta-studies in many dimensions, so the studies are difficult to compare.

4.2 Ad Effectiveness Funnel Elasticity

We wish to relate the relative incremental lift in upper-funnel outcomes to the relative incremental lift in lower-funnel outcomes. The marketing funnel describes the consumer journey from awareness to purchase. The funnel is a metaphor for the attrition that arises as consumers move through the stages along the path to purchase. Here, we consider site visits as our upper-funnel outcome and conversions as our lower-funnel outcome. This relationship will help practitioners and academics use the upper-funnel lift estimates to evaluate or optimize ad campaigns when lower-funnel outcomes are unavailable or too noisy.

Practitioners and academics often lack the sales and profit data from the bottom of the marketing funnel to determine the return on investment of advertising. When such data is available, Lewis & Rao (2015) point out that sales data is highly variable which hurts the statistical power of lift estimates. Instead, practitioners and academics frequently rely on upper funnel outcomes to evaluate advertising effectiveness. Many online display ad field experiments in the literature measure ad effects using various upper funnel outcomes: survey metrics including purchase intent and favorability (see e.g. Goldfarb & Tucker 2011; Bart

[3]Note that the binomial tests of $p \leq 0.2$ and $p \leq 0.015$ of our data provide the same qualitative result.

et al. 2014), online search (e.g. Lewis & Nguyen 2015), site visits (e.g. Hoban & Bucklin (2015)), and likes on Facebook (e.g. Bakshy et al. 2012). The challenge lies in how marketers translate the effect of an ad on upper funnel outcomes to the subsequent effect on sales. If an ad generates a 10% lift in site visits, does this translate into less than, greater than, or equal to a 10% lift in sales? Firms could better optimize their ad choices if firms understood this relationship, yet previous work implicitly assumed but did not quantify this relationship. Here, we seek to quantify the relationship between the causal effect of advertising between outcomes in the marketing funnel. Though the relationship between this upper and lower funnel lifts varies across advertisers, we explore the distribution of this ratio and posit that marketing academics and practitioners could find the mean/median useful as an initial rule-of-thumb.

In Figure 3, we distinguish between the baseline and the incremental marketing funnel. The baseline funnel describes the attrition along the user's path to purchase in the absence of the focal campaign. Here, the baseline funnel is identified by the control group users' site visit y_C^U and conversion y_C^L outcomes. The incremental funnel describes the attrition along the purchase funnel for those incremental outcomes that are caused by the ads. The incremental funnel is identified by the experimental difference in user site visits, $y_T^U - y_C^U$, and conversions, $y_T^L - y_C^L$,between the treatment and control groups. The conversion rate in the baseline funnel is $r_C \equiv y_C^L/y_C^U$ whereas the conversion rate in the incremental funnel is $r_\Delta \equiv \left(y_T^L - y_C^L\right) / \left(y_T^U - y_C^U\right)$.

The elasticity of the ad effectiveness funnel (denoted by ε_∇) relates the proportional lift in the upper funnel outcome to the proportional lift in the lower funnel outcome. That is,

$$\varepsilon_\nabla \equiv \frac{y_\Delta^L}{y_\Delta^U} = \frac{y_T^L - y_C^L}{y_C^L} / \frac{y_T^U - y_C^U}{y_C^U}. \tag{7}$$

Figure 3 illustrates this relationship. The elasticity can also be interpreted as the ratio of the baseline conversion rate and the incremental conversion rate, as $\varepsilon_\nabla = \frac{y_T^L - y_C^L}{y_T^U - y_C^U} / \frac{y_C^L}{y_C^U} = \frac{r_\Delta}{r_C}$.

Note that ε_∇ is an elasticity in that it relates the percent change in the two lifts. However, ε_∇ does not refer to a lever that the firm can pull as is the case with the price or advertising elasticity of demand. Nonetheless, ε_∇ can be thought of as the ratio between elasticities: the elasticity of advertising with respect to site visits and with respect to conversions in our case. When ε_∇ is elastic ($\varepsilon_\nabla > 1$), then the incremental site visits convert more often than those in the baseline funnel. When ε_∇ is inelastic ($\varepsilon_\nabla < 1$), then the incremental site visits convert less often than those in the baseline funnel. Since both the baseline and incremental users who visit the advertiser's website are self-selected, whether ε_∇ will be elastic or inelastic is unclear. Indeed, the literature provides conflicting evidence for the ad effectiveness funnel elasticity: online display ads in one field experiment were inelastic (Johnson et al., 2016a) and search ads were inelastic in one setting (Dai & Luca, 2016) but elastic in another (Sahni, 2015). Since the interpretation of ε_∇ depends on the direction that it differs from 1, we seek to reject the null hypothesis $H_0 : \varepsilon_\nabla = 1$ by running a t-test.

The heterogeneity in outcome data requires that we filter the studies to meaningfully estimate ε_∇. First, we limit our analysis to the 92 experiments for which we have distinct site visit and conversion outcome data and for which site visitors exceed converters. All the studies satisfying this first step satisfy our power condition from Section 3.1. Second, we require that the lift in the upper-funnel outcome satisfy the significance threshold that the t-statistic > 1. Until now, we avoided restricting the studies based on their estimated lift in order to ensure that the results are representative. Since y^U_Δ is the the denominator in eq. (7), y^U_Δ close to 0 will make ε_∇ very large and give it undue influence in the average. Moreover, requiring y^U_Δ to be positive ensures that ε_∇ will have the same sign as y^L_Δ, so that the sign of ε_∇ is easier to interpret. Requiring $t > 1$ is analogous to passing a one-sided test with $\alpha = 0.159$, a slightly stricter criterion than the meta-analysis in Lodish et al. (1995a) which uses a one-sided $\alpha = 0.20$. These restrictions for both site visitors and visits reduce our sample to 69 experiments.

We find that the ad effectiveness funnel is most often inelastic. Table 2 presents our

average estimates for ε_∇ and t-tests for $\varepsilon_\nabla = 1$ for different interquantile ranges of ε_∇ from 100% down to 80% to avoid influential outliers. Examining all studies, we find that the average $\varepsilon_\nabla = -0.967$ ($t = -0.05$) when comparing incremental site visitors to incremental converters. As the data window narrows, the average ε_∇ becomes significantly different from 0 with $\varepsilon_\nabla = 0.694$ (t=-2.30) for the 95% interquantile range and $\varepsilon_\nabla = 0.569$ (t=-5.57) for the 80% interquantile range. Similarly, we find the average $\varepsilon_\nabla = 0.568$ (t=-0.57) when comparing incremental visits and conversions for all studies. As the data window narrows, the average ε_∇ again significantly deviates from $\varepsilon_\nabla = 1$ with $\varepsilon_\nabla = 0.671$ (t=-1.88) for the 95% interquantile range and $\varepsilon_\nabla = 0.596$ (t=-4.75) for the 80% interquantile range. ε_∇ is thus consistently between 0.6 and 0.7 when we ignore outliers. In other words, an advertiser with a 10% lift in site visits should expect on average to find a 6% or 7% lift in conversions.

Figures 4 and 5 illustrate the distribution of the funnel elasticity separately at the extensive and overall margins. We see that $\varepsilon_\nabla < 1$ the majority of the time; however, we see heterogeneity in the average funnel elasticity even within the 95% interquantile range of elasticities and after filtering out many experiments. Table 3 summarizes the heterogeneous distribution of elasticities for the individual tests: the median of ε_∇ is 0.51 for visitors to converters and 0.58 for visits to conversions.

4.3 Carryover Effects

We measure the carryover effect of online display advertising after the campaigns end. To begin, we denote the cumulative outcome $y(\tau)$ from the beginning of the campaign to τ days after the end of the campaign. The cumulative *absolute* lift estimator after τ days is then

$$\Delta y(\tau) \equiv E\left[y(\tau) | Z = T, \hat{D} = 1\right] - E\left[y(\tau) | Z = C, \hat{D} = 1\right]$$

for the default PGA estimator. The fallback ITT estimator only conditions on treatment assignment (see eq. 4). The cumulative nature of the outcome means that the cumulative

absolute lift estimator $\Delta y(\tau) = \Delta y(0) + (\Delta y(\tau) - \Delta y(0))$ can be decomposed into the in-campaign absolute lift $\Delta y(0)$ and absolute carryover $\Delta y(\tau) - \Delta y(0)$, which is the post-campaign lift. We define the *relative carryover*, which normalizes the absolute carryover by the in-campaign absolute lift, as

$$Carryover_\tau \equiv \frac{\Delta y(\tau) - \Delta y(0)}{\Delta y(0)} = \frac{\Delta y(\tau)}{\Delta y(0)} - 1. \tag{8}$$

Thus, $Carryover_\tau > 0$ would be in line with most analytical models which posit a positive carryover (see e.g. Nerlove & Arrow 1962); whereas $Carryover_\tau < 0$ means that the ads cause users to substitute their online site visits forward in time.

Estimating the carryover effect of a campaign presents a statistical power problem. Lewis & Rao (2015) draw attention to the statistical power challenge in measuring online display ad effectiveness: ad effects are small relative to the variance in marketing outcomes, so that ad estimates will be imprecise even in large samples. Lewis et al. (2015) elaborate that this problem is compounded for measuring carryover because the absolute value of the carryover effect is plausibly smaller than the during-campaign effect—for a given time period. In other words, the lift in the outcome is smaller after the campaign but the noise in the outcome variable is unchanged after the campaign. Consequently, we expect our carryover estimates to be highly variable and should be interpreted accordingly. Given that single studies are imprecise, we leverage the meta-study to draw general conclusions about the form of carryover.

Due to the statistical power problem, we focus on the outcomes with greater effect size and more study observations: site visits and site visitors. We also restrict our sample in several ways. We begin with the 339 tests that meet out power threshold (see Section 3). To eliminate deactivated pixels from the data, we drop cases where the number of visits and visitors does not change in a week for either the treatment or control groups. We measure the campaign lift up to 28 days after the campaign, so we only examine the 128 experiments in

our sample that include at least 28 days of post-experiment data. To improve our statistical power, we restrict our analysis to estimates with $t > 1$ for $y_\Delta(0)$. Again, this is analogous to passing a one-sided test with $p = 0.159$ following Lodish et al. (1995b) who use a one-sided $p = 0.20$ criterion. Also, restricting attention to positive in-campaign lift estimates also allows for a straightforward interpretation of the relative carryover: the relative carryover will have the same sign as the absolute carryover. We impose the $t > 1$ restriction on both the site visits (drops 36 studies) and visitors (drops 29 studies) since the two have substantial overlap (39 studies). These restrictions reduce our sample to 89 studies.

We are interested in both the sign and magnitude of the relative carryover. In Figures 6 and 7, we see the distribution of the relative carryover 28 days after the campaign for both site visitors and site visits. Throughout, we restrict attention to the 83 studies in the 95% interquantile range of the $Carryover_\tau$ variable to avoid influential outliers. The relative carryover estimates are heterogeneous and skew towards the positive side for visits, but the visitor carryover estimates are balanced between positive and negative carryover. Among all 89 studies, the relative carryover is weakly positive 49% of the time for visitors and 65% of the time for site visits. Like in Section 4.2, we test the null hypothesis that carryover is zero using a t-test. Table 4 presents the average carry-over for both visitors and visits and for 7, 14, 21 and 28 days after the campaign. All estimates suggest a positive carryover that is highly significant for visits (t-statistics$>$ 2.88) but less so for visitors (t-statistics between 1.18 and 2.00). After 7 days, the average carryover is 2.6% (t=1.93) for visitors and 6.3% (t=3.05) for visits. After 28 days, the average carryover is 6.2% (t=2.00) for visitors and 16.4% (t=2.88) for visits. Table 5 presents the percentiles of the relative carryover across weeks. We see that the median carryover estimates are more modest: 0.0% for visitors and 2.9% for visits 28 days after the campaigns. The median advertiser should therefore expect little or no carryover after the campaign.

Our relative carryover estimates suggest a more modest effect than the split-cable TV studies in the second meta-study by Lodish et al. (1995b). They focus on 42 studies that

demonstrate a significant lift during the campaign at the one-sided $\alpha = 0.2$ level. They estimate an average of about a 100% relative carryover. Nonetheless, they consider a longer time horizon: the 2-year post campaign lift from a 1-year television campaign. Moreover, cookie churn will attenuate our carryover estimates as users who delete their cookies can no longer be connected to their later outcomes on the advertiser's site. Still, Sahni (2015) suggests much higher carryover effect in search ads for restaurants: the ad effect decays exponentially at a rate of about 0.90 per week. Though our estimates are noisy, they suggest that some campaigns may have had a negative carryover. In future research, we hope to explore the determinants of the sign and magnitude of the carryover as well as appropriate functional form assumptions for the carryover effect.

5 Conclusion

In this paper, we present strong evidence that online display ads increase site visits and conversions. While we see heterogeneity in relative ad effectiveness throughout our study, the median lift is 16% for visits and 8% for conversions. We also find heterogeneity when we measure the elasticity of the ad effectiveness funnel and the carryover effect. Nonetheless, we suggest rules of thumb for marketers: incremental upper-funnel lift on average mean a less-than-proportional lower-funnel lift (elasticity of roughly 0.5-0.7) and the average four-week carryover effect is 6% for visitors and 16% for visits. These rules of thumb should help marketers make better use of their ad effectiveness estimates.

This meta-analysis provides a big-picture view of ad effectiveness measurements. However, a key theme here is the heterogeneity in ad effects and relationships that caution against a single theory. This heterogeneity is a natural consequence of different advertisers from different industries, choosing different advertising goals, and measuring different outcomes. This suggests that individual advertisers will find repeated experimentation helpful for identifying regularities in the effectiveness of their own ads.

References

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).

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*.

Bleier, A. & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, 34(5).

Clarke, D. G. (1976). Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research*, (pp. 345–357).

Coey, D. & Bailey, M. (2016). People and cookies: Imperfect treatment assignment in online experiments. In *WWW 2016*.

Dai, W. & Luca, M. (2016). Effectiveness of paid search advertising: Experimental evidence. In *Presented at NBER Economics of IT and Digitization 2016*.

eMarketer (2016). US digital display ad spending to surpass search ad spending in 2016.

Facebook (2016). How we're making ad measurement more insightful. Facebook for Business News.

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

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

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

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2016). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. White Paper.

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.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467–475.

Johnson, G. A., Lewis, R. A., & Nubbemeyer, E. I. (2016a). Ghost ads: Improving the economics of measuring ad effectiveness. *Available at SSRN*.

Johnson, G. A., Lewis, R. A., & Reiley, D. (2016b). When less is more: Data and power in advertising experiments. *Marketing Science*. Forthcoming.

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

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. A. (2010). *Measuring the Effects of Online Advertising on Human Behavior Using Natural and Field Experiments*. PhD thesis, MIT Dept of Economics.

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

Lewis, R. A. & Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics*, 130(4), 1941–1973.

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

Lodish, L. M., Abraham, M. M., Livelsberger, J., Lubetkin, B., Richardson, B., & Stevens, M. E. (1995b). A summary of fifty-five in-market experimental estimates of the long-term effect of TV advertising. *Marketing Science*, 14(3), G133–G140.

Nerlove, M. & Arrow, K. J. (1962). Optimal advertising policy under dynamic conditions. *Economica*, 29(114), pp. 129–142.

Sahni, N. (2015). Effect of temporal spacing between advertising exposures: Evidence from an online field experiment. *Quantitative Marketing and Economics*, 13(3), 203–247.

Sahni, N. (2016). Advertising spillovers: Field experimental evidence and implications for returns from advertising. *Journal of Marketing Research*, 53(4).

Sethuraman, R., Tellis, G. J., & Briesch, R. A. (2011). How well does advertising work? generalizations from meta-analysis of brand advertising elasticities. *Journal of Marketing Research*, 48(3), 457 – 471.

Shrivastava, A. (2015). Understanding the impact of Twitter ads through conversion lift reports. Twitter blog.

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

Figures & Tables

Figure 1: Incremental site visits across all 347 experiments with 95% confidence intervals

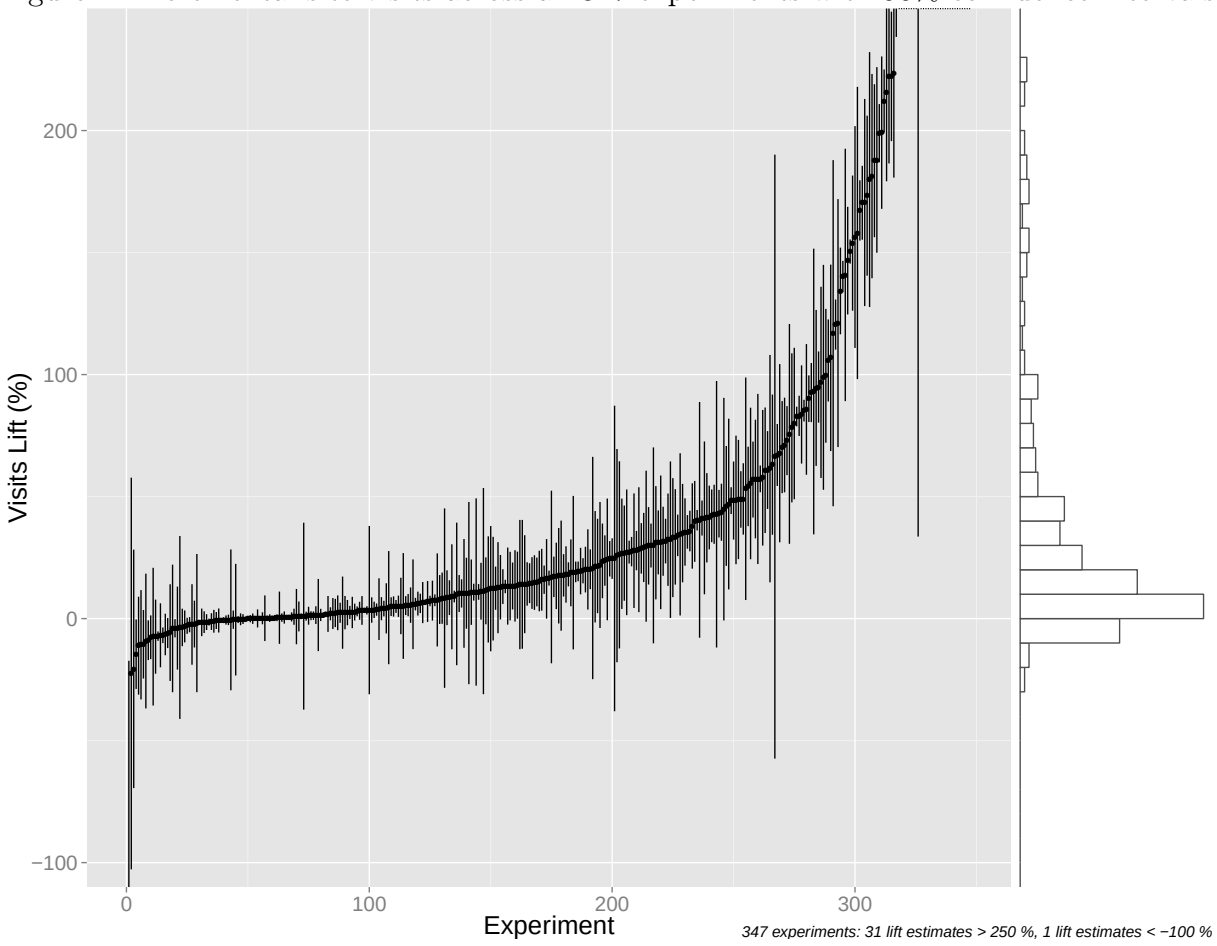


Figure 2: Incremental conversions across all 184 experiments with 95% confidence intervals

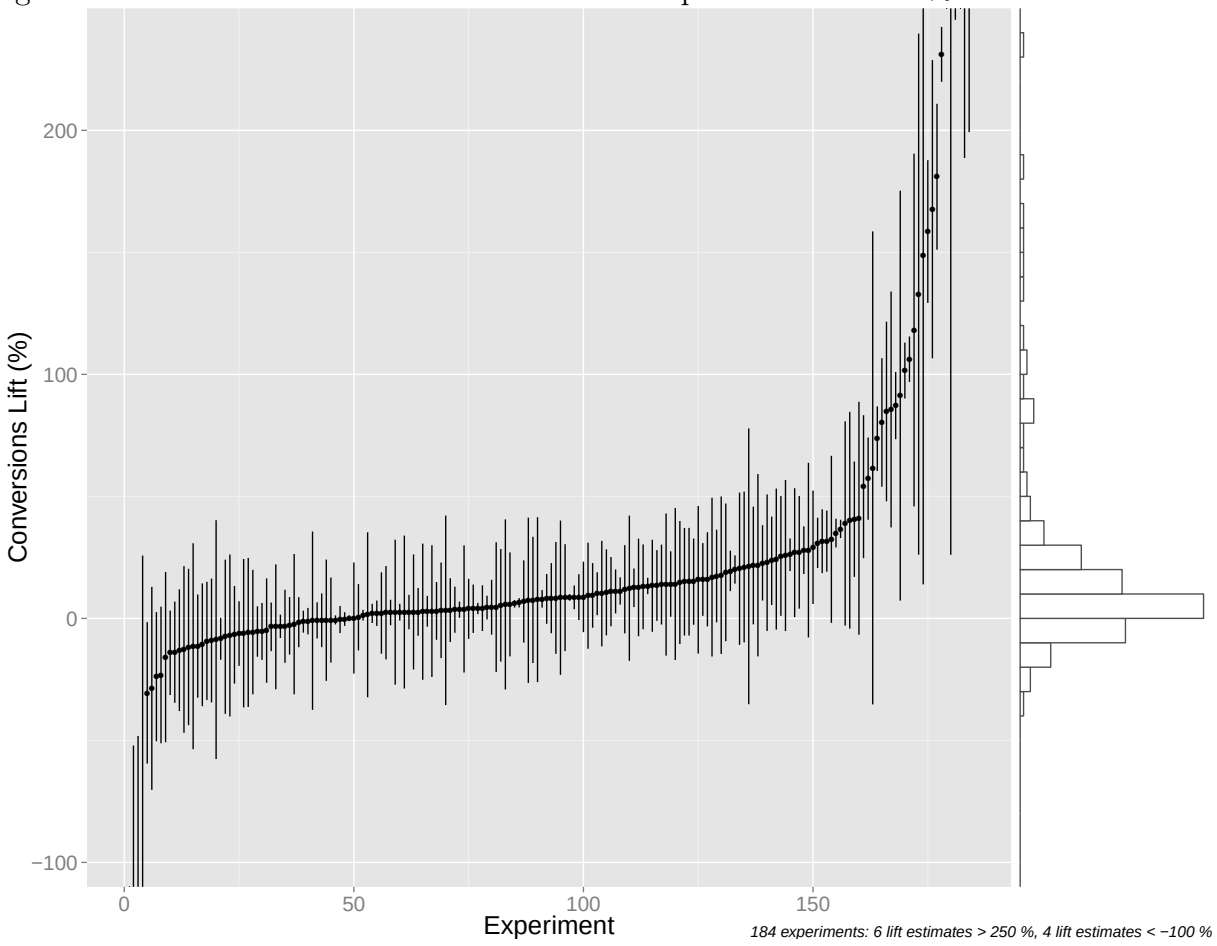


Figure 3: Ad effectiveness funnel elasticity: Relating lower funnel to upper funnel outcomes

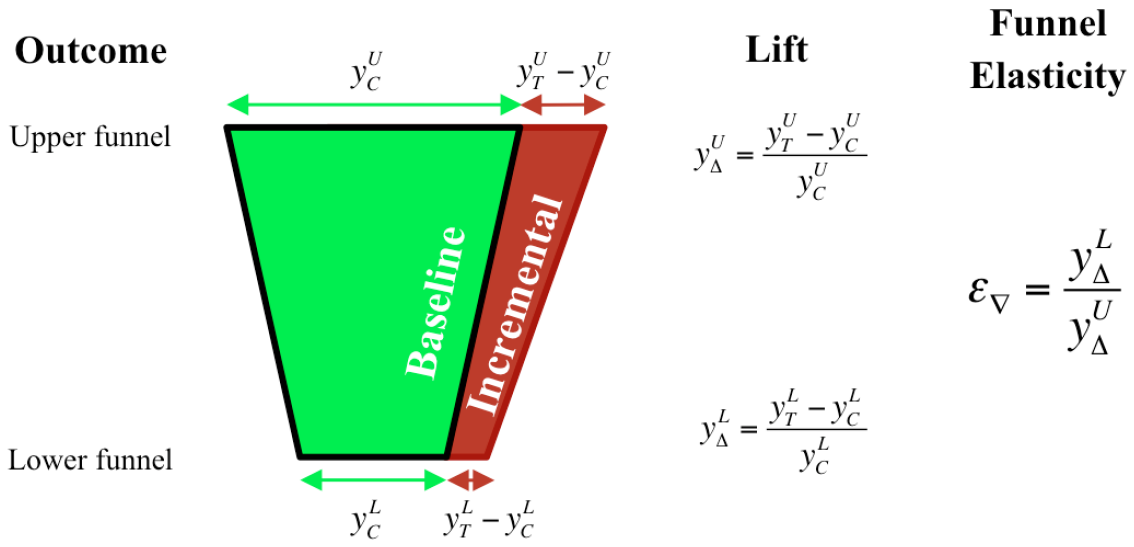
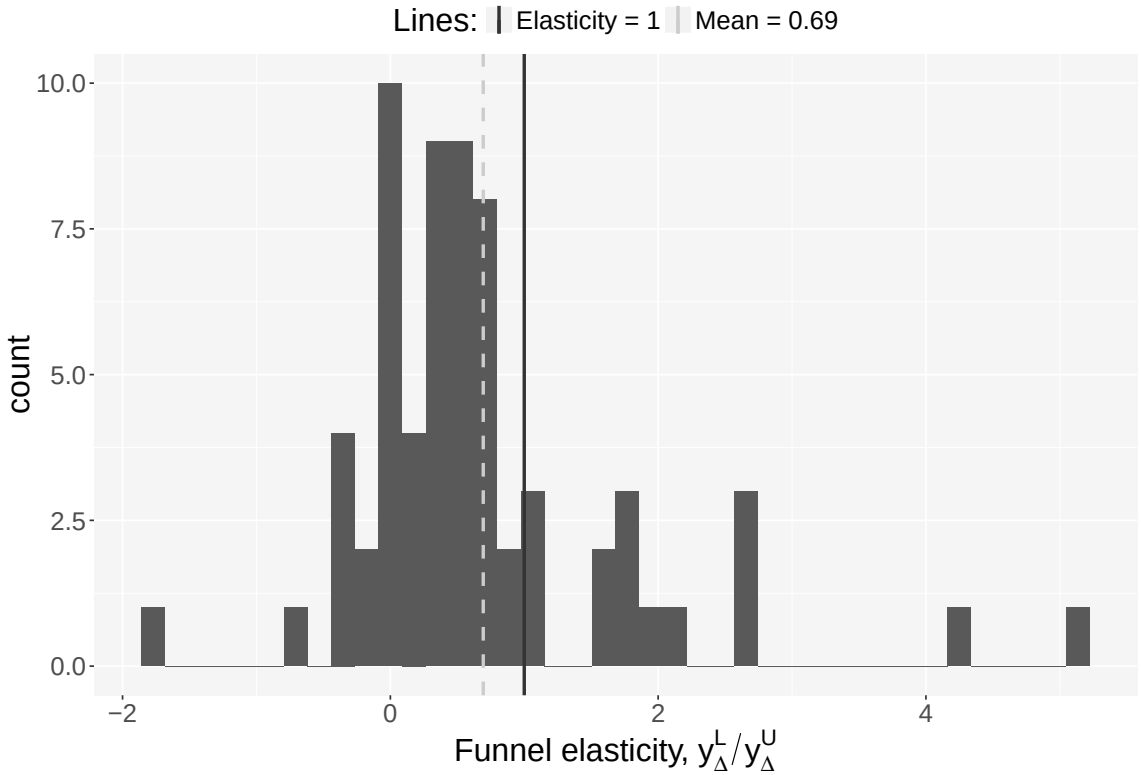
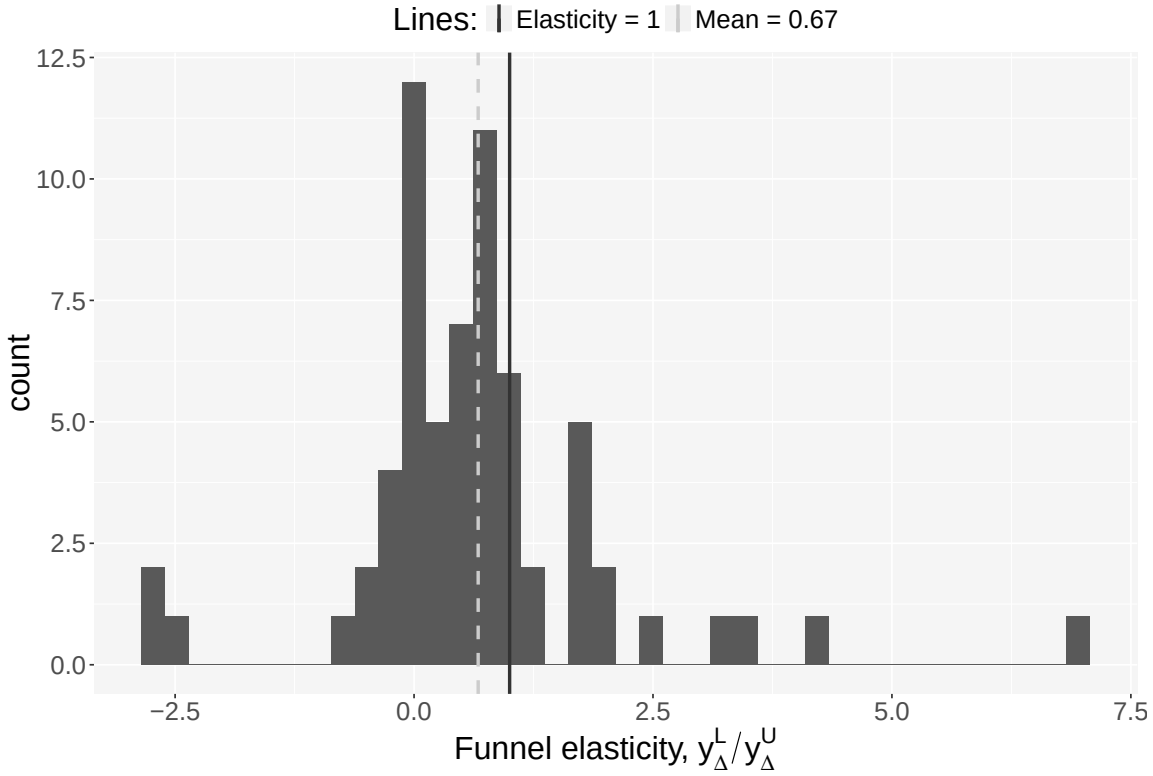


Figure 4: Extensive ad effectiveness funnel: Incremental visitors to converters



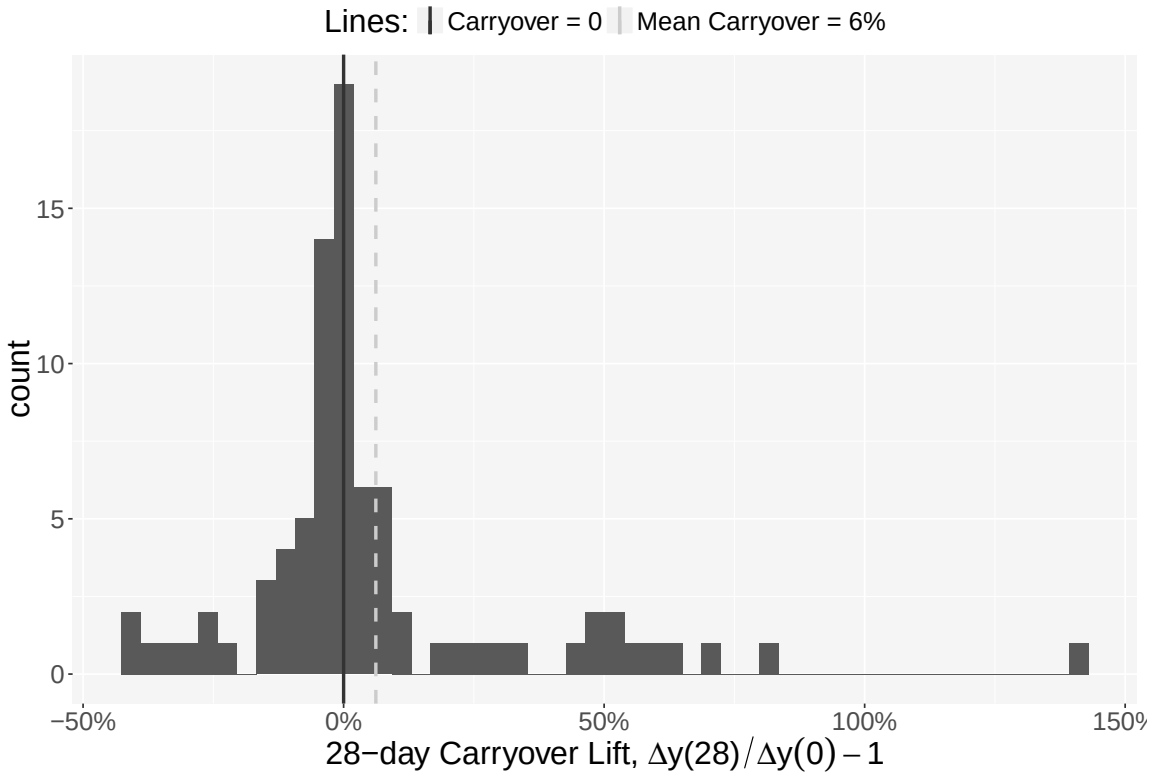
Notes: Includes 65 studies in the 95% interquantile range of the funnel elasticity.

Figure 5: Ad effectiveness funnel: Incremental visits to conversions



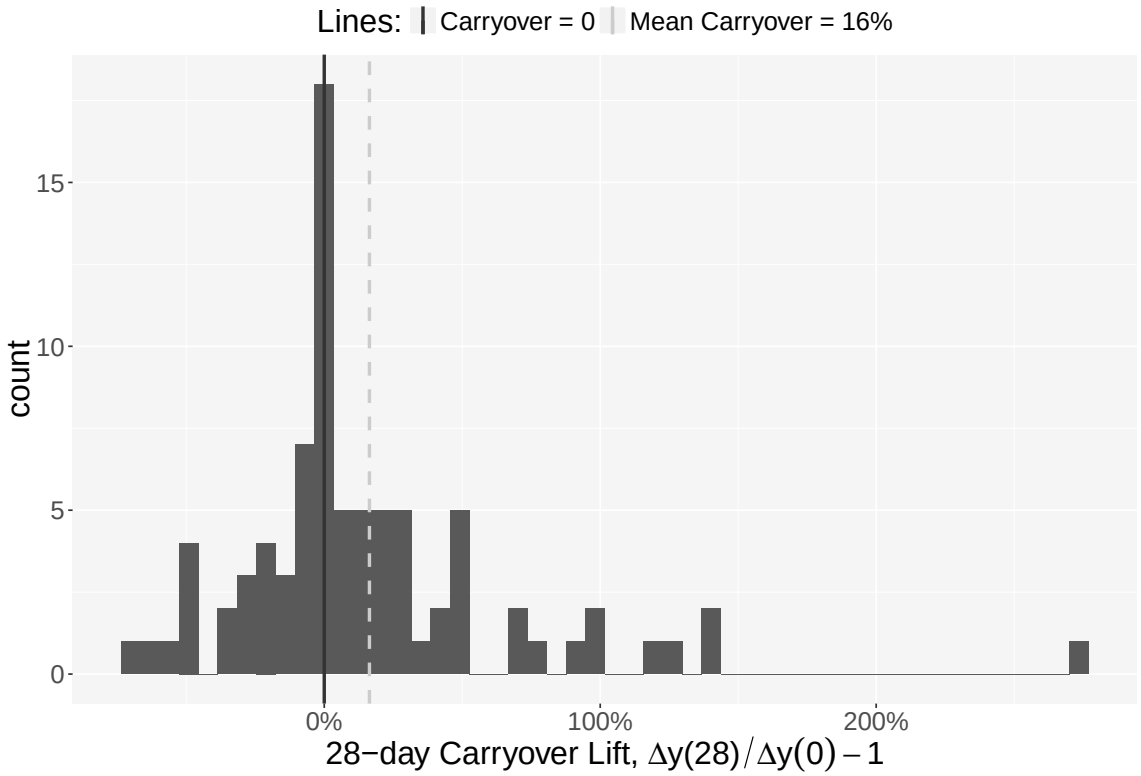
Notes: Includes 65 studies in the 95% interquantile range of the funnel elasticity.

Figure 6: Relative Carryover: Visitors



Notes: Includes 83 studies in the 95% interquantile range of the outcome $\frac{\Delta y(28)}{\Delta y(0)}$.

Figure 7: Relative carryover: visits



Notes: Includes 83 studies in the 95% interquantile range of the outcome $\frac{\Delta y(28)}{\Delta y(0)}$.

Table 1: Overall lift across 432 ghost ad experiments

	All Experiments†				Powerful Experiments†			
	Visits	Visitors	Conversions	Converters	Visits	Visitors	Conversions	Converters
Median lift estimate	16.6%	21.8%	8.1%	7.9%	15.9%	21.2%	8.0%	7.7%
[.10, .90]-quantile	[-1.1%, 213.6%]	[-0.2%, 284.4%]	[-8.9%, 83.4%]	[-7.2%, 67.5%]	[-1.1%, 180.2%]	[-0.2%, 203.8%]	[-7.4%, 50.1%]	[-6.5%, 54%]
[.025, .975]-quantile	[-8.1%, 780.9%]	[-4.8%, 830.3%]	[-29.5%, 258.8%]	[-26.2%, 252%]	[-7.5%, 559.3%]	[-4.5%, 569.4%]	[-15.3%, 176.7%]	[-12.8%, 184.5%]
Average lift estimate	1189.2%	699.7%	19.9%	19.8%	75.1%	88.7%	20.9%	22.5%
Standard error	(1098.9%)	(596.1%)	(10.8%)	(8.5%)	(11.1%)	(12.2%)	(3.8%)	(4.2%)
Weighted average lift estimate††	836.4%	534.9%	25.3%	34.4%	112.6%	133.0%	32.1%	40.7%
Standard error	(705.7%)	(384.1%)	(14.6%)	(20%)	(29.8%)	(36.7%)	(14.5%)	(20.7%)
Individual significance tests								
Reject 5% two-sided	202	249	57	65	194	243	48	57
Reject 2.5% one-sided (Lift <0)	7	8	4	5	6	8	1	1
Reject 2.5% one-sided (Lift >0)	195	241	53	60	188	235	47	56
Collective significance†††	7.38E-213	1.43E-296	2.93E-40	6.11E-49	9.35E-204	6.52E-289	1.79E-34	1.98E-45
N. of Experiments	347	347	184	184	339	339	174	174

Notes: †All experiments includes all experiments with a given outcome for which we observe at least 100 expected control-group users engaging in the outcome. Powerful experiments filters out those tests whose standard errors are too large to detect a 100% lift in incremental users over the control group baseline 95% percent of the time (st. err $*3.3 < y_C$). ††Lift estimates are weighted by the number of treatment-ad exposed users in the test. †††The collective significance test uses a binomial test for the realized number of studies with positive lift that reject a 2.5% one-sided lift versus a null of 2.5%.

Table 2: Average ad effectiveness funnel elasticity

Visitors to Converters				
Interquantile range	Elasticity (ε_∇)	St. Err.	t of $H_0 : \varepsilon_\nabla = 1$	Studies
80%	0.569	0.077	-5.57	55
90%	0.627	0.096	-3.88	61
95%	0.694	0.133	-2.30	65
100%	0.967	0.724	-0.05	69

Visits to Conversions				
Interquantile range	Elasticity (ε_∇)	St. Err.	t of $H_0 : \varepsilon_\nabla = 1$	Studies
80%	0.596	0.085	-4.75	55
90%	0.622	0.119	-3.18	61
95%	0.671	0.175	-1.88	65
100%	0.568	0.756	-0.57	69

Notes: Includes at most 69 powerful studies with both visit & conversion outcomes & for which t-stats on the upper funnel lift>1 for both visits and visitors.

Table 3: Distribution of ad effectiveness funnel elasticity

Visitors to Converters						
Quantiles of Elasticity (ε_∇)						
Min	10%	25%	50%	75%	90%	Max.
-26.743	-0.367	0.061	0.509	0.880	2.199	31.547

Visits to Conversions						
Quantiles of Elasticity (ε_∇)						
Min	10%	25%	50%	75%	90%	Max.
-40.197	-0.515	0.045	0.580	1.064	2.148	26.272

Notes: Includes 69 powerful studies with both visit & conversion outcomes & for which t-stats on the upper funnel lift>1 for both visits and visitors.

Table 4: Average relative carryover

Visitors				Visits			
τ (days)	$\frac{\Delta y(\tau)}{\Delta y(0)} - 100\%$	Std. Err.	t-statistic	τ (days)	$\frac{\Delta y(\tau)}{\Delta y(0)} - 100\%$	Std. Err.	t-statistic
7	2.6%	1.3%	1.93	7	6.3%	2.1%	3.05
14	2.0%	1.7%	1.18	14	8.9%	2.8%	3.21
21	5.5%	2.7%	2.00	21	13.9%	4.4%	3.16
28	6.2%	3.1%	2.00	28	16.4%	5.7%	2.88

Notes: We include the powerful studies for which we observe 28 days of activity post-campaign and for which t-stats>1 for both visits and visitors. To avoid influential outliers, we only examine the 83 of 89 studies in the 95% interquantile range of the outcome $\frac{\Delta y(\tau)}{\Delta y(0)}$.

Table 5: Distribution of relative carryover

Visitors:	Relative Carryover Quantiles				
τ (days)	10%	25%	50%	75%	90%
7	-8.3%	-1.7%	0.5%	6.8%	20.8%
14	-18.6%	-3.3%	0.8%	7.3%	24.9%
21	-17.6%	-5.1%	0.3%	8.8%	34.5%
28	-24.4%	-5.9%	0.0%	8.6%	52.6%

Visits:	Relative Carryover Percentiles				
τ (days)	10%	25%	50%	75%	90%
7	-20.0%	-1.5%	3.6%	12.9%	34.2%
14	-23.1%	-3.2%	4.7%	22.2%	54.4%
21	-34.4%	-5.7%	3.9%	27.2%	76.6%
28	-46.9%	-8.6%	2.9%	30.7%	97.1%

Notes: We include the 89 powerful studies for which we observe 28 days of activity post-campaign and for which t-stats>1 for both visits and visitors.

Appendix

A ITT Estimator

As we discuss in Section 2.3, Intent-to-treat (ITT) requires that we can delineate users who are eligible for exposure and track outcomes among eligible users. In some field experiments, the subjects are a list of users defined prior to the campaign (see e.g. the database match campaign in Johnson et al., 2016b). Here, no such list exists. To resolve this, we use our deterministic user randomization algorithm to sort all the advertiser's site visitors into treatment and control eligible users. Effectively, we define eligibility as all N Internet users who could visit the site and sum the $N^{\emptyset}$ such users with non-zero outcomes, which suffices to measure the sum $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N^{\emptyset}} y_i^{\emptyset}$. Given that the experiments all assign 70% of users to the treatment group and 30% to control, our ITT estimator computes the total campaign effect as

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

Without knowing N, we cannot compute the exact standard errors of the total ITT estimator. We instead use a conservative approximation of the variance $Var(\sum_{i=1}^{N} y_i) \approx \sum_{i=1}^{N^{\emptyset}} \left(y_i^{\emptyset}\right)^2$ following Appendix B of Johnson et al. (2016a) who note that

$$Var(\sum_{i=1}^{N} y_i) = \sum_{i=1}^{N} y_i^2 - \frac{1}{N} \left(\sum_{i=1}^{N} y_i \right)^2 \leq \sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N^{\emptyset}} \left(y_i^{\emptyset}\right)^2$$

One challenge in meta-studies is to normalize the estimated lifts across studies in such a way that they are comparable. This challenge is magnified when we use two different treatment effect estimators. For comparability, we normalize the test's lift estimates to be the relative lift over the baseline outcomes in the control group rather than the absolute lift. However, while the total lifts between the LATE_{PGA} and ITT estimates are the same in expectation whenever PGA does not under-predict treated users (see eq. 6), the baseline

total outcome in ITT will be higher than LATE_{PGA} because ITT outcomes include outcomes from users who would not see a focal ad. Consequently, the relative lift for ITT estimates would be unfairly low relative to LATE_{PGA}. To remedy this, we use the baseline outcomes among the predicted-exposed control group users for both the LATE_{PGA} and ITT estimates. Since the Hausman test rejections indicate that the PGA predicted-exposed control group sample may not match the treatment group sample, this approach may bias our results. We examine the ITT studies by hand to eliminate any obviously flawed studies. Otherwise, we think our approach picks the best among imperfect options.