The Online Display Ad Effectiveness Funnel & Carryover: Lessons from 432 Field Experiments

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

October 1, 2017

Abstract

We analyze 432 online display ad field experiments on the Google Display Network. The experiments feature 431 advertisers from varied industries, which on average include 4 million users. Causal estimates from 2.2 billion observations help overcome the medium's measurement challenges to inform how and how much these ads work. We find that the campaigns increase site visits ($p < 10^{-212}$) and conversions ($p < 10^{-39}$) with median lifts of 17% and 8% respectively. We examine 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. We then 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.

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

*Johnson: Kellogg School of Management (visiting), Northwestern University, <garrett.johnson@kellogg.northwestern.edu>. Lewis: Netflix, <randall@econinformatics.com>. Nubbemeyer: Google, <elmarn@google.com>. We thank Abdelhamid Abdou, David Broockman, Hubert Chen, Jennifer Cutler, 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). In 2016, online display supplanted search as the largest contributor to the largest ad spending category: digital advertising. Much remains to be learned about the effectiveness of online display advertising. However, the medium's effects are tiny relative to the variation in marketing outcomes, which poses two measurement challenges. First, observational methods fail to recover experimental estimates (Gordon et al., 2017), so experiments represent the gold standard in measurement (Lavrakas, 2010). Second, ad effectiveness measurement suffers from an extreme statistical power problem, which limits what can be learned from individual campaigns (Lewis & Rao, 2015). The median field experiment in Lewis & Rao (2015) requires 3.3 million exposed consumer observations to differentiate between break-even and ineffective campaigns, and 1.3 billion observations to detect a 5% difference in marketing outcomes. These challenges pose a quandary for marketers: experiments are critical for learning ad effectiveness, yet little can be learned due the power problem.

We use a unique data set of 2.2 billion observations across 432 field experiments to explore the effectiveness of online display advertising. The experimental ad campaigns include 431 advertisers across varied industries all using the Google Display Network (GDN). GDN is one of the largest online display ad platforms, encompassing 2 million websites and reaching over 90% of global users (Google, 2015). On average, these experiments reach over 4 million users and last 20 days. Google did not share with us *any* data about the advertisers, campaigns, creatives, or exposed users beyond the experimental lift in order to protect advertiser's privacy. We begin by providing novel, broad-based evidence that online display ads increase user site visits and conversions for the advertiser—by a median of 17% and 8% respectively. We find modest carryover in ad effectiveness four weeks after the campaign. Finally, we find that incremental changes in a proxy outcome (site visits) overstate incremental changes in user conversions.

Recent studies by Blake et al. (2015) and Shapiro (2016) engender skepticism about the effectiveness of advertising, as they document cases where advertisers spent millions of dollars without finding statistically or economically significant ad effects. We demonstrate with exceptionally strong statistical evidence that online display advertising changes consumer behavior online. We observe when users visit the advertiser's website for 347 studies. The median lift in visits is 16.6%, but the 90%-interquantile range of [-1.1%, 213.6%] reveals much dispersion. 195 of these 347 studies demonstrate a significantly positive lift ($p = .025$, one-sided). A binomial collective significance test rejects at the $p < 10^{-212}$ level that 195 of these 347 studies are significant by random chance alone. Moreover, we show that the strength of this collective evidence exceeds that of past meta-studies of ad field experiments—regardless of media or outcomes. We observe conversion outcome data in 184 studies, with a median lift of 8.1% and a 90%-interquantile range of [-8.9%, 83.4%]. 53 of these 184 studies demonstrate significantly positive lift, which is collectively significant at the $p < 10^{-39}$ level by the same criterion.

We further leverage our data to address two challenges that marketers face. First, many marketers evaluate a campaign's effectiveness based on its short-run performance during the campaign alone. After all, the long run effect is not yet observed when managers end a campaign, and the statistical power problem is even worse for measuring long run rather than short run effects (Lewis et al., 2015). However, this myopic approach ignores any effect of the ads—positive or negative—that carries over after the campaign ends. If these carryover effects are significant and positive, a myopic marketer would undervalue their previous campaigns and underinvest in future campaigns. Second, many firms lack data that connect advertising exposures to purchase outcomes, and so they instead rely on proxy metrics such as survey responses, clicks, ad engagements, online search volume, or site visits. Marketers often employ a rule-of-thumb to translate the relative lift in a proxy metric into a corresponding change in purchases. However, if the users who incrementally increase the proxy metric are less (more) likely to purchase than baseline users, then the marketer

would over-(under-)invest in advertising. Below, we evaluate and propose models for these challenges.

Our data enable us to study the carryover effect of advertising after the campaign ends. The dominant model of carryover in marketing is the Koyck or ad-stock model (see e.g. Nerlove & Arrow 1962). This model posits that the sign of the carryover effect is positive and that the magnitude decays geometrically over time. Formally, the effect of Ad_t j periods later is given by $\delta^j \cdot Ad_{t-j}$ for some $\delta > 0$. Much past work takes this assumption for granted. This assumption underpins time series work on ad effectiveness, which employs distributed lag model regressions for marketing outcomes on advertising (see Sethuraman et al. 2011 for a summary). We critically examine the ad stock model by evaluating the model's fit, its generalizability and its implied carryover parameter (δ). This extends efforts using field experiments to test the sign (Lodish et al., 1995b; Simester et al., 2009) and functional form (Sahni, 2015) implications of the ad stock model.

We measure the economic importance of the carryover effect and use the time path of carryover estimates to evaluate the ad stock model. We find that carryover is modest four week after the average campaign. For site visits, the lift increases on average a further 16.4% above the lift during the campaign and a further 2.9% on median. On the extensive margin, the average carryover for visitors is 6.2%, but the median campaign has no carryover (0.0%). When we examine the time path of the carryover estimates across campaigns, three broad categories emerge. 58% of campaigns exhibit positive carryover in visits. A geometric model improves fit over a linear model in half of those cases. When the geometric model performs significantly better, the median carryover coefficient is $\delta = 0.77$ per day. In 32% of cases, the carryover is negative—falling by a median of 1.2% daily. Lastly, 10% of campaigns show no significant evidence of non-zero carryover. Thus, substantial heterogeneity in carryover exists, which requires both richer ad models and marketer-specific learning.

Our study of proxy outcomes builds on the canonical concept of the marketing funnel (Strong, 1925). The marketing funnel describes the attrition as consumers move through

stages along the path to purchase. Our related notion of the *ad effectiveness funnel* describes the incremental effect of ads at different stages of the purchase funnel. We expect an incremental lift in upper funnel outcomes to translate into some incremental lift in lower funnel outcomes. A crucial but under-explored question is: how much? For instance, consider a campaign that causes 0.1% of exposed users to visit a website where 20% of site visitors typically convert. In the absence of conversion data, a manager may assume that 0.02%=0.1%*20% of exposed users will incrementally convert. Whether this assumption is reasonable hinges on whether incremental visits are more or less likely to convert than baseline visits. Understanding how incremental effects of ads propagate down the marketing funnel is useful for imputing bottom line outcomes when marketers lack this information. Much academic research too relies on upper funnel metrics to measure marketing success. Furthermore, when multiple outcomes are present, we can learn more from a campaign by pooling information while leveraging the relationship between incremental outcomes.

We provide a first look at the ad effectiveness funnel using a large number of experiments with both upper funnel (i.e., site visit) and lower funnel (i.e., conversion) outcomes. We propose a simple proportional relationship between the relative lifts in the lower and upper funnel outcomes. We find that incremental site visits are less likely to convert than baseline site visits. On median, a 10% lift in visits leads to a 5.1% lift in conversions and a 10% lift in visitors leads to an 5.8% lift in converters. Indeed, the relative lift in site visits exceeds the relative lift in conversions in over 70% of the campaigns. This means that most marketers—when assuming a proportional lift throughout the marketing funnel—would overinvest in advertising.

Meta-studies play an important role within the larger marketing literature on how advertising works (see e.g. Vakratsas & Ambler 1999 for a summary). Meta-studies provide a wide view of causal ad effect estimates, and pool information across studies to overcome the statistical power problem. Table 1 summarizes meta-studies of ad field experiments in the marketing literature. These meta-studies cover three ad media: television (e.g. Lodish et al.,

1995a), online search (e.g. Kalyanam et al., 2015), and online display. Among online display ad studies, the largest meta-studies demonstrate the medium's effectiveness for stated preference outcomes (Goldfarb & Tucker, 2011a; Bart et al., 2014). We provide broad-based evidence that online display ads lift site visits and conversions, which had been limited to a search platform (Sahni, 2015) and Facebook (Gordon et al., 2017). Though some meta-studies in Table 1 include more studies (Goldfarb & Tucker, 2011a) or larger average sample sizes (Gordon et al., 2017), the present meta-study has the largest subject-study combination (2.16 billion users across 432 studies) of all those we review.

The ad stock model has received limited scrutiny in ad field experiments. Lodish et al. (1995b) find positive carryover two years only after successful year-long television ad campaigns, which is directionally consistent with an ad stock model. Zantedeschi et al. (2014) applies the ad stock model to field experiment panel data to show that catalogs have longer lasting effects than emails. However, Sahni (2015) shows an effect of ad spacing, which suggests a richer carryover model founded on psychological theories of memory. Finally, Simester et al. (2009) provide an example of negative carryover in catalog advertising, which suggests an alternative mechanism: advertising causes consumers to purchase earlier than they would have otherwise. Our contribution undermines the generalizability of the ad stock model as few studies exhibit a positive, geometrically decaying carryover and many campaigns appear to only shift user visits forward in time. These efforts contribute to an emerging literature that tests the functional forms implied by marketing theory using ad field experiments (see also Lewis, 2014; Sahni, 2016).

To the best of our knowledge, we are the first to explore the relationship between incremental lifts in multiple marketing funnel outcomes across a large number of field experiments. Past work in this spirit uses proxy measures like survey outcomes to predict baseline purchase behavior (Dyson et al., 1996). This relationship between baseline outcomes is then extrapolated to conclude that display ads lift sales based on incremental survey evidence alone (Briggs, 1997). Other researchers instead document heterogeneous ad effects by user's

funnel stage prior to the experiment (e.g. Hoban & Bucklin, 2015). Broad-based field experiments that include multiple funnel outcomes to date only exist for paid search advertising (Sahni, 2015; Dai & Luca, 2016). We make novel use our study-level lift estimates to model the relationship between incremental funnel outcomes. Our incremental funnel and our carry-over estimates can then serve as rules of thumb to help marketers evaluate their campaigns with limited data.

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 carryover estimates, and the ad effectiveness funnel elasticity estimates. Section 5 concludes.

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. (2016).

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. (2016) 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. (2016) 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. (2016) 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. (2016) 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. (2016), 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. 2016). 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 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. (2016) 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

[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.

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. (2017), 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. (2017) 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.

4 Meta-Analysis Results

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

4.1 Overall Treatment Effects

Recent case studies call into question the effectiveness of advertising (Blake et al., 2015; Shapiro, 2016), arguing that naive observational estimates may lead managers to overestimate the true effect of advertising. Most prior evidence showing that online display ads work is arises from stated rather than revealed preference outcomes (Goldfarb & Tucker, 2011b; Bart et al., 2014). We begin by presenting the the ad effectiveness estimates from 432 tests, for which we observe site visits for 347 tests and conversions for 184 tests.

[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., 2016), which makes for a weaker test.

We normalize each test's lift by the baseline outcome, which is established by the control group. This helps compare across tests that vary in many dimensions including the campaign's duration and 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 2 summarizes the lift estimates 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 the reach-weighted average are high at about 1200% and 800% due to influential outliers[3]. 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.

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 2 shows, the overall

[3]Within the 339 powerful studies (defined in Section 3), the unweighted and weighted average lifts plummet to 75% and 113%.

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 online consumer behavior. We use a binomial test to evaluate the joint statistical significance of the findings. Table 2 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. The binomial test returns a p-value of $7.4{*}10^{-213}$ for visits and $p = 2.93{*}10^{-40}$ for conversions. The extensive margin evidence is even stronger with, a p-value of $p = 1.43{*}10^{-296}$ for visitors and $6.11 * 10^{-49}$ for conversions.

Returning to Table 1, we see that the collective significance of our findings compares favorably with prior meta-studies of ad field experiments. The Lodish et al. (1995a) split cable television experiments find that only 115 (39%) of the 292 weight tests are significant for sales at the $p \leq 0.2$ one-sided level. An analogous binomial test in the last column of Table 1 shows the collective significance of these results is high at $p = 2.35{*}10^{-14}$. Follow-up studies by Hu et al. (2007, 2009) are bolstered by including matched-market tests yielding collective significance of $p = 4.31 * 10^{-62}$ and $p = 8.25 * 10^{-21}$ respectively. To be clear, these comparisons do not mean that online display ads 'work' better than TV ads, rather our lift estimates are more precisely estimated because our sample sizes average 4 million users rather than under 20,000 households. In online display ads, we compute analogous joint significance statistics of $p = 1.99 * 10^{-13}$ for purchase intention in Bart et al. (2014), and $p = 2.55 * 10^{-13}$ for conversions in Gordon et al. (2017). Our collective evidence for the lift in site visits exceeds all the ad studies we review in Table 1 and the evidence for conversions exceeds all but the Hu et al. (2007) television studies.

This strong evidence is somewhat surprising given that our sample is constructed in such

a way that the outcomes we examine do not necessarily correspond with the objectives of the campaign. In some cases, the objective was brand-building measured by brand favorability in user surveys, so the lift in site visits and conversions was incidental. One potential concern is that non-human traffic (bots) can bias online outcomes. However, the conversion evidence mitigates this concern in that bots do not trigger many conversion outcomes like making purchases. Non-human traffic is also a small portion of online users though may contain outliers with many site visits; thus, our extensive margin evidence for visitors and converters further mitigates the concern that bots drive these results.

4.2 Carryover Effects

We measure the carryover effect of online display advertising after the campaigns end. Our data has a combination of features, which enables a novel exploration of online display ad carry-over: its importance to the median firm, its variance across advertiser campaigns and its evolution over time. First, we observe an end date for for the experiment after which additional exposure—if any—should be identical across campaigns. We then directly observe the carryover effect without needing to separate this from the contemporaneous ad effect as in past work (e.g. Dubé et al. 2005). Second, our data have a daily panel structure, which reveals how carryover evolves. While past work often imposes the geometric decay assumption of the ad stock/goodwill model, our panel data provide a novel opportunity to test this assumption. Third, some studies are statistically powerful, which ameliorates the ad carryover measurement challenge. Fourth, our large collection of studies provides a broad view of carryover.

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

$$Carryover_{\tau} \equiv \frac{\Delta y(\tau) - \Delta y(0)}{\Delta y(0)} = \frac{\Delta y(\tau)}{\Delta y(0)} - 1. \tag{7}$$

Thus, $Carryover_{\tau} > 0$ would be in line with ad stock models (see e.g. Nerlove & Arrow 1962) which posit a positive carryover; 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) 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, our carryover estimates contain much sampling variation. We therefore pool across studies to better understand 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. Next, we require that

the in-campaign lift $\Delta y\,(0)$ 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 $\Delta y\,(0)$ is the the denominator in eq. (7), $\Delta y\,(0)$ close to 0 will make $Carryover_\tau$ very large and give it undue influence in the average. 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. Requiring $t > 1$ is analogous to passing a one-sided test with $\alpha = 0.159$, stricter than the $\alpha = 0.20$ criterion used in Lodish et al. (1995a). 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 3 and 4, we see the distribution of the relative carryover 28 days after the campaign for both site visitors and site visits. 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 4.3, we test the null hypothesis that carryover is zero using a t-test. Table 3 presents the average carryover for both visitors and visits and for 7, 14, 21 and 28 days after the campaign. We restrict attention to the 83 studies in the 95% interquantile range of the $Carryover_\tau$ variable to avoid influential outliers. 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 4 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. The upper and lower quantiles reveal carryover of opposite sign that

diverges over time. The 10% quantile for the relative carryover for visitors falls from -8.3% after one week to -24.4% after four weeks; whereas, the 90% quantile rises from 20.8% to 52.6% over that period. The relative carryover in visits diverges even more: -20.0% after one week to -46.9% after four weeks in the 10% quantile and 34.2% to 97.1% in the 90% quantile. In other words, the during campaign lift is twice the total lift inclusive of carryover for the 10% quantile and half the total lift for the 90% quantile.

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. Many differences could explain the differences in performance. For instance, online display ads may receive less attention from consumers and the medium may be less conducive to brand-building. 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, which we explore using panel data on the carryover effect below.

Carry-over time path [preliminary]

We leverage our data's daily panel structure to examine the evolution of carryover. We show the time path of each campaign's carryover, which reveals persistent patters obscured in the cross-campaign data above. We then use this data to test the dominant goodwill or ad stock model in marketing, which assumes ad carryover is positive and decays geometrically.

We begin by non-parametrically estimating the time path of carryover. Figure 5a and 5b illustrate the evolution of each campaign's cumulative carry-over for visits and visitors respectively during the 28 days after the campaign. These are daily lift estimates that we do not smooth. The confidence intervals on these estimates show both the challenge of

measurement precision but also the opportunity that the most statistically powerful studies provide. The confidence intervals for the visitor estimates are narrower because that outcome is binary, and is less variable. Despite the noise, many studies demonstrate either persistent positive (e.g. study #439) or negative (e.g. #676) trends.[4] In general, a study's visitor trend follows its visit trend, though visitor trends are more negative. This happens because the visitor outcome by definition reduces the pool of potential new visitors over time among (predicted) exposed users. Study #268 is then notable in that its visit and visitor trends move in opposite directions: an additional lift of about 20% in visits but -10% in visitors after four weeks. The persistent trends exhibited by several studies suggest that the cross-campaign carryover heterogeneity observed in Figure 4 reflects fundamental differences in the sign and magnitude of carryover by advertiser rather than sampling variation alone.

We sort the studies into three groups based on whether their time-series data exhibit significantly positive, negative, or neither trend in cumulative carryover. To do so, we fit the cumulative carryover data in Figure 5 to linear models and test the null hypothesis of zero carryover trend at the 5%, two-sided level. We then sort the campaigns by visit and visitor carryover as follows:

Linear trend	Visits	Visitors
Positive, significant	58%	38%
Negative, significant	32%	45%
Insignificant	10%	17%

The goodwill model posits a positive carryover, so it cannot explain the 32% of studies with negative trends for site visits. Even more visitor estimates are negative again because the pool of new visitors shrinks over time. The 38% of studies with positive trends in visitors thus suggest that some campaigns can have important lasting effects.

[4] A few studies (notably #175) appear to show no initial carry-over followed by a sharp increase in ad lift. These patters are surprising and we are gathering additional data to determine whether the advertiser began a second experiment that was not recorded in our initial data. Until then, we include all studies in the interest of completeness and transparency.

Finally, we fit our carryover estimates to a goodwill model of advertising. We condition on the 58% of studies with positive linear trends in visits as we expect the goodwill model will perform poorly otherwise. We fit the panel data of daily post-campaign lift estimates (not cumulative) to the ad stock model below:

$$\frac{E\left[y_{i\tau}|\widehat{D_i}=1,Z_i=T\right]-E\left[y_{i\tau}|\widehat{D_i}=1,Z_i=C\right]}{\Pr\left[D_i=1|\widehat{D_i}=1,Z_i=T\right]}=\beta_i\delta_i^{\tau}+\varepsilon_{i\tau}$$

for advertiser campaign i and post-campaign day τ. β_i represents the lift from the cumulative ad stock as campaign i ends, which depreciates at rate δ_i on subsequent days. Note that β_i does not equal the in-campaign lift or the additional lift on the last day of the campaign as these include both the carryover effect as well as the contemporaneous effect of advertising. As such, β_i is estimated from the post-campaign lift data alone.

Even among positive studies, a linear model often improves fit over an ad-stock model.[5] For visits, the ad-stock model performs better 56% of the time based on the Akaike information criterion and 48% of the time based on the Bayesian information criterion. In those cases, the median estimates for δ_i are 0.84 and 0.77 respectively. These rates imply that 90% of the cumulative carryover effect occurs in the first 13.1 and 8.8 days respectively.

Discussion

Our results suggest that the carryover effects of online display advertising should be thought of as the net effect of competing positive (goodwill) effects and negative (inter-temporal substitution) effects. Moreover, advertisers vary considerably by which force dominates. However, most past literature only emphasizes the positive carry-over effects of advertising. In particular, the ad-stock/goodwill model is broadly used in marketing as an identifying assumption to estimate the effect of advertising from observational data using distributed

[5]The gradual decay of the ad stock model is realistic and simplifies the long run ad effect calculation. The linear model does not share these features, but it is simple. In the future, we intend to consider a wider class of carryover models (Huang et al., 2012).

lag models (Sethuraman et al., 2011 reviews 56 such studies). This approach is now applied in online display advertising (e.g. Braun & Moe, 2013; Bruce et al., 2017). Our novel data provide new evidence that the ad stock model is limited in its generalizability for online display advertising.

Some caveats apply to our carryover results. 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. Our data is limited to the experiments the advertisers ran in this time period. Some advertisers may have run subsequent non-experimental campaign and any cross-campaign interaction effects would create differences between the treatment and control group beyond the carryover. The direction of this bias would depend on whether the experimental and non-experimental campaigns are net complements of substitutes in the second period.

4.3 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 and the attrition that arises as consumers move through the stages along this journey. Here, we consider site visits as our upper-funnel outcome and conversions as our lower-funnel outcome. We explore this relationship because it will help practitioners and academics use the upper-funnel lift estimates to learn from campaigns when lower-funnel outcomes are unavailable or too noisy.

In Figure 6, 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,

$$\varepsilon_\nabla \equiv \frac{y_\Delta^L}{y_\Delta^U} = \frac{y_T^L - y_C^L}{y_C^L} / \frac{y_T^U - y_C^U}{y_C^U}. \tag{8}$$

Figure 6 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: 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 as before satisfy the significance threshold

that the t-statistic > 1. 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. 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. 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 5 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 7 and 8 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 6 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.

More research is needed to extend this finding to more ad media and more funnel outcomes. Still, this finding calls into question the assumption that incremental effect on the funnel outcomes is proportional and suggests a more conservative managerial takeaway from upper funnel lift outcomes. The interpretation of $\varepsilon_\nabla < 1$ is that incremental users are less

likely to progress down the funnel. Our evidence in Table 2 that the median lift in visitors (21.8%) exceeds the lift in visits (16.6%) supports this interpretation as it means incremental visitors are less active on the advertiser's website. Still, a more mundane explanation could be that incremental site visitors include bots, which are less likely to convert. Alternately, $\varepsilon_{\nabla} < 1$ could result from a potential mismatch between the goal of the campaign and our conversion outcome.

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 17% for visits and 8% for conversions. We also find heterogeneity when we measure the carryover effect and the elasticity of the ad effectiveness funnel. Nonetheless, we suggest rules of thumb for online display ad campaigns: the average four-week carryover effect is 6% for visitors and 16% for visits and the incremental upper-funnel lift on average mean a less-than-proportional lower-funnel lift (elasticity of roughly 0.5-0.7). 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

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*, 51(3), 270–285.

Blair, M. H. (1987). An empirical investigation of advertising wearin and wearout. *Journal of Advertising Research*, 40(06), 95–100.

Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, 83(1), 155–174.

Bleier, A. & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, 34(5), 669–688.

Braun, M. & Moe, W. W. (2013). Online display advertising: Modeling the effects of multiple creatives and individual impression histories. *Marketing Science*, 32(5), 753–767.

Briggs, R. (1997). *IAB Online Advertising Effectiveness Study*. Technical report, Internet Advertising Bureau.

Bruce, N. I., Murthi, B., & Rao, R. C. (2017). A dynamic model for digital advertising: The effects of creative format, message content, and targeting on engagement. *Journal of Marketing Research*, 54(2), 202–218.

Coey, D. & Bailey, M. (2016). People and cookies: Imperfect treatment assignment in online experiments. In *Proceedings of the 25th International Conference on World Wide Web*.

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

Dubé, J.-P., Hitsch, G. J., & Manchanda, P. (2005). An empirical model of advertising dynamics. *Quantitative marketing and economics*, 3(2), 107–144.

Dyson, P., Farr, A., & Hollis, N. S. (1996). Understanding, measuring, and using brand equity. *Journal of Advertising Research*, 36(6), 9–22.

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

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.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2017). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook.

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.

Hu, Y., Lodish, L. M., Krieger, A. M., & Hayati, B. (2009). An update of real-world tv advertising tests. *Journal of Advertising Research*, 49(2), 201–206.

Huang, J., Leng, M., & Liang, L. (2012). Recent developments in dynamic advertising research. *European Journal of Operational Research*, 220(3), 591 – 609.

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. (2016). Ghost ads: Improving the economics of measuring ad effectiveness. *Available at SSRN*.

Johnson, G. A., Lewis, R. A., & Reiley, D. (2017). When less is more: Data and power in advertising experiments. *Marketing Science*, 36(1), 43–53.

Kalyanam, K., McAteer, J., Marek, 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. Available at https://ssrn.com/abstract=2684110.

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., 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. (2014). Worn-out or just getting started? The impact of frequency in online display advertising. Working Paper.

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.

MacInnis, D. J., Rao, A. G., & Weiss, A. M. (2002). Assessing when increased media weight of real-world advertisements helps sales. *Journal of Marketing Research*, 39(4), 391–407.

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: Evidence from online field experiments and implications for returns on advertising. *Journal of Marketing Research*, 53(4), 459–478.

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.

Shapiro, B. (2016). Advertising in health insurance markets. *Available at SSRN*.

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

Strong, E. K. (1925). *The psychology of selling and advertising*. McGraw-Hill book Company, Incorporated.

Vakratsas, D. & Ambler, T. (1999). How advertising works: what do we really know? *The Journal of Marketing*, 63(1), 26–43.

Zantedeschi, D., Feit, E. M., & Bradlow, E. T. (2014). Measuring multi-channel advertising response using consumer-level data.

Figures & Tables

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

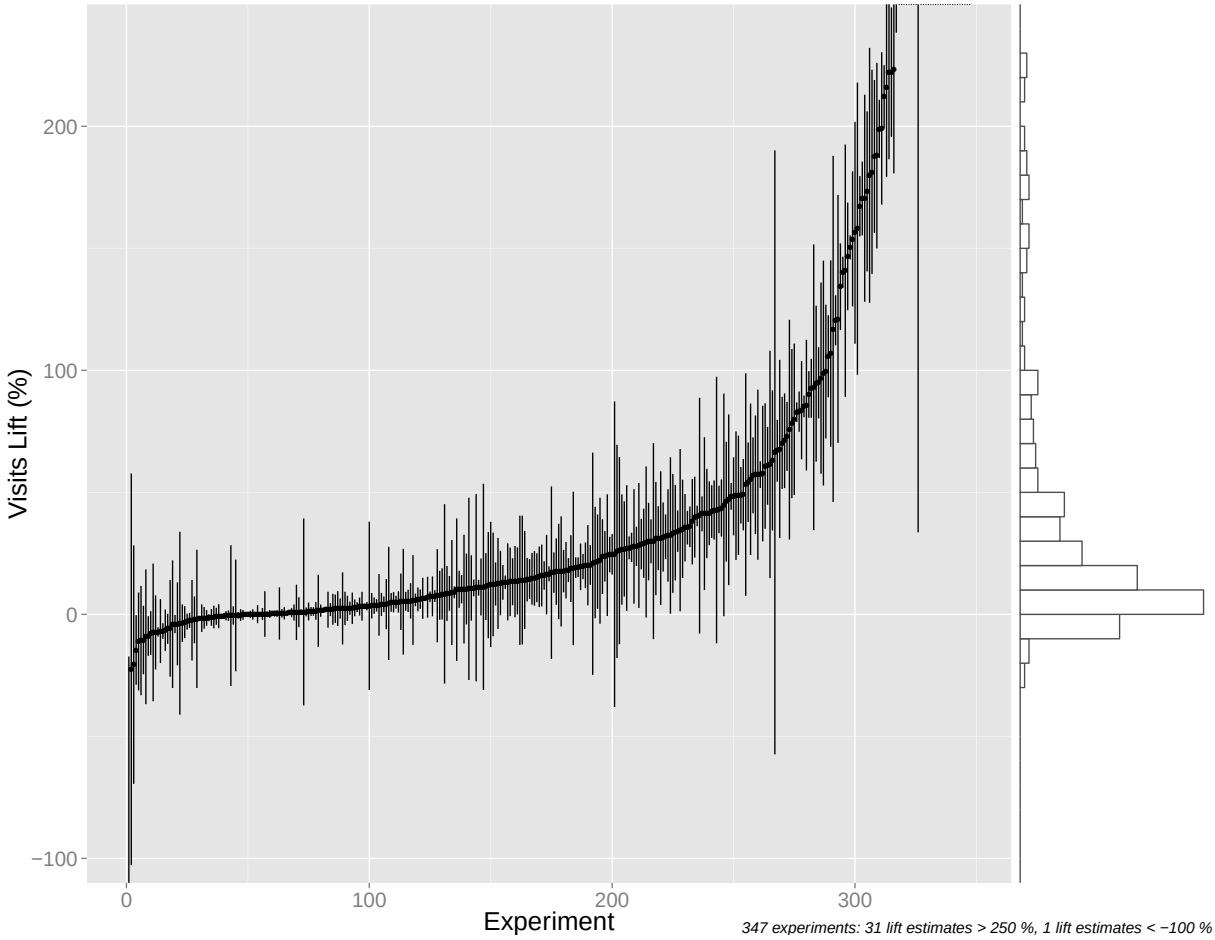


Figure 2: Incremental conversions across all 184 experiments with 95% confidence intervals

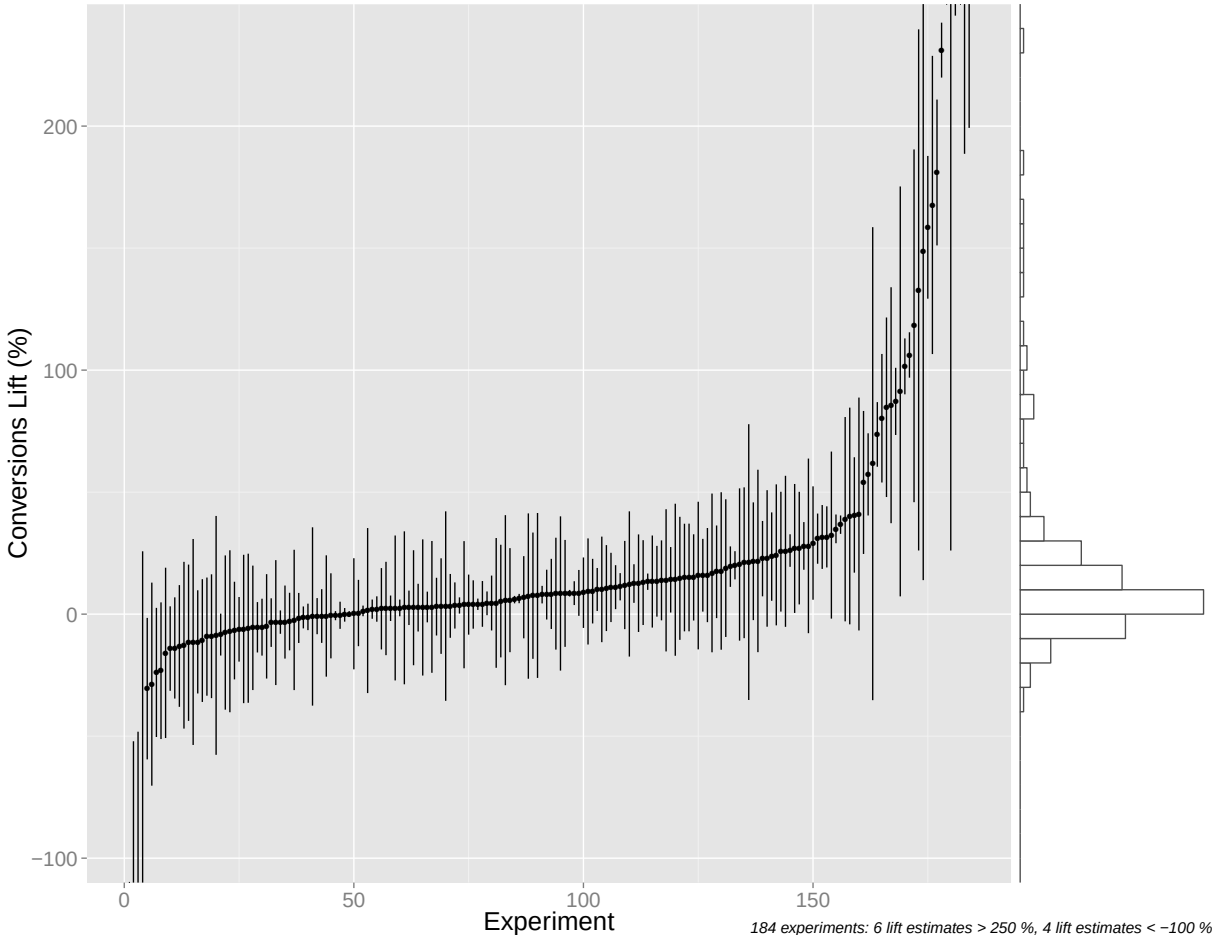
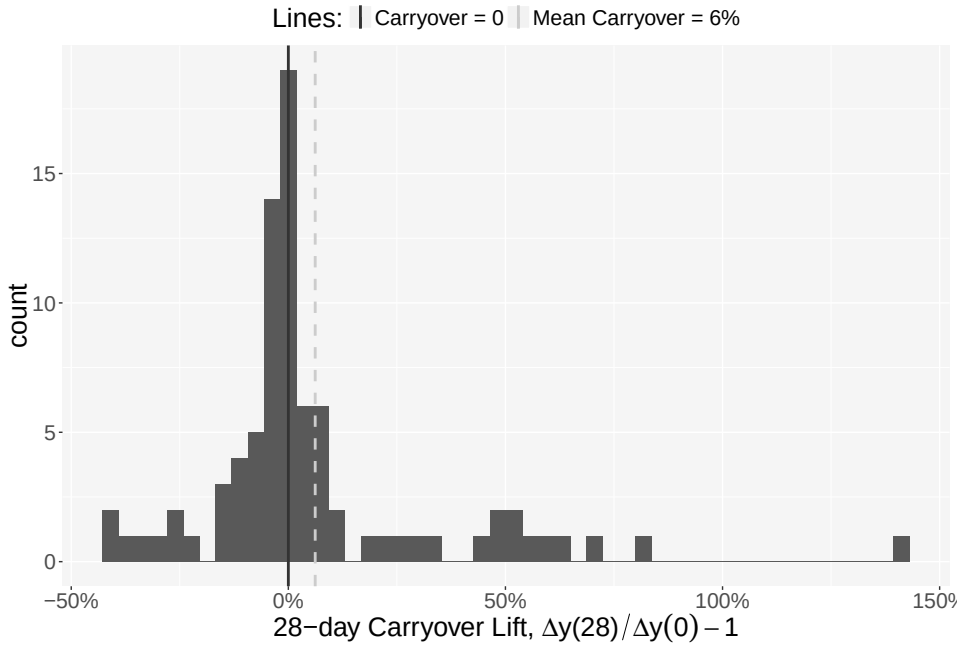
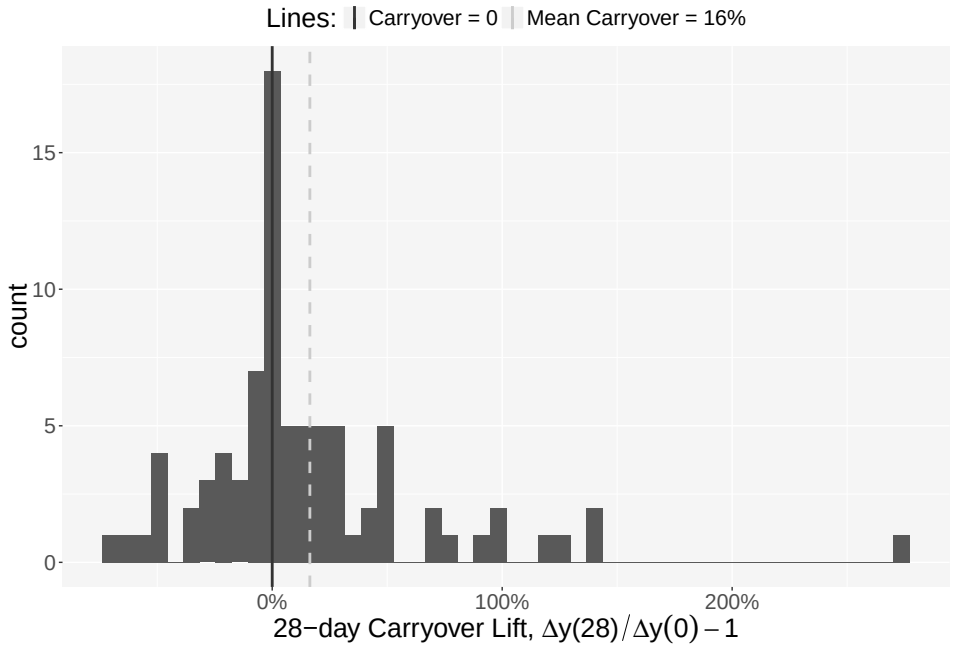


Figure 3: Relative carryover: visitors



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

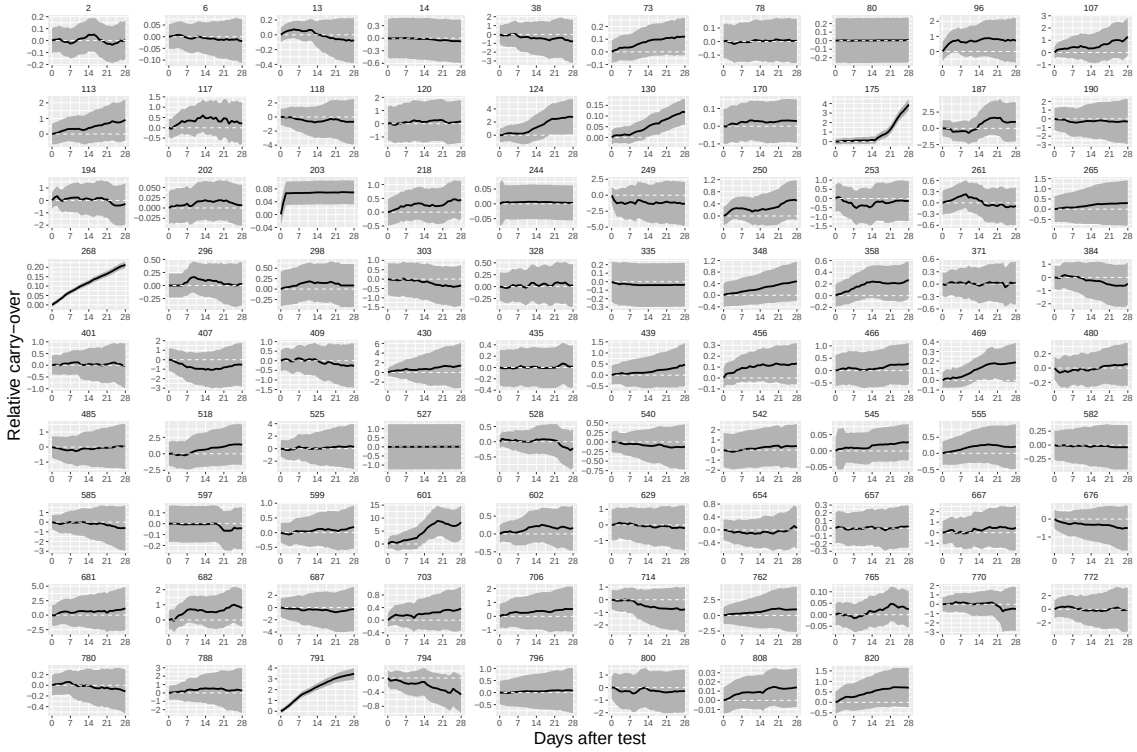
Figure 4: Relative carryover: visits



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

Figure 5: Time path of cumulative carryover by campaign

(a) Site visits



(b) Site visitors

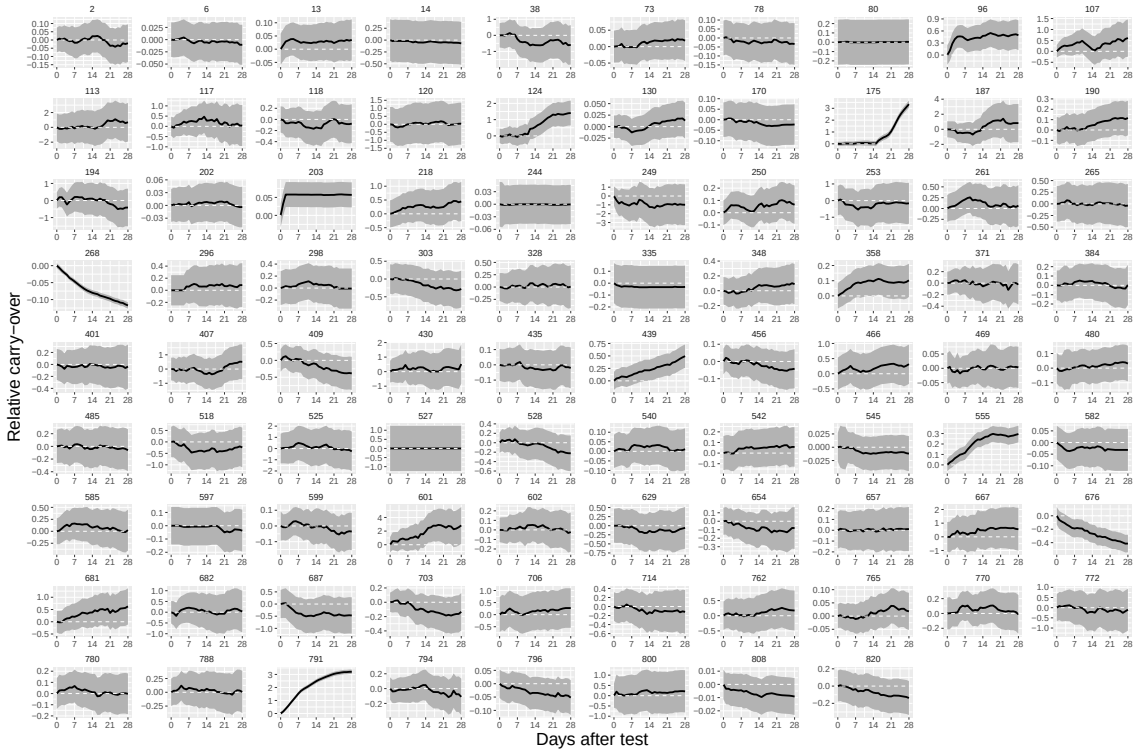


Figure 6: Ad effectiveness funnel elasticity: Relating lower funnel to upper funnel outcomes

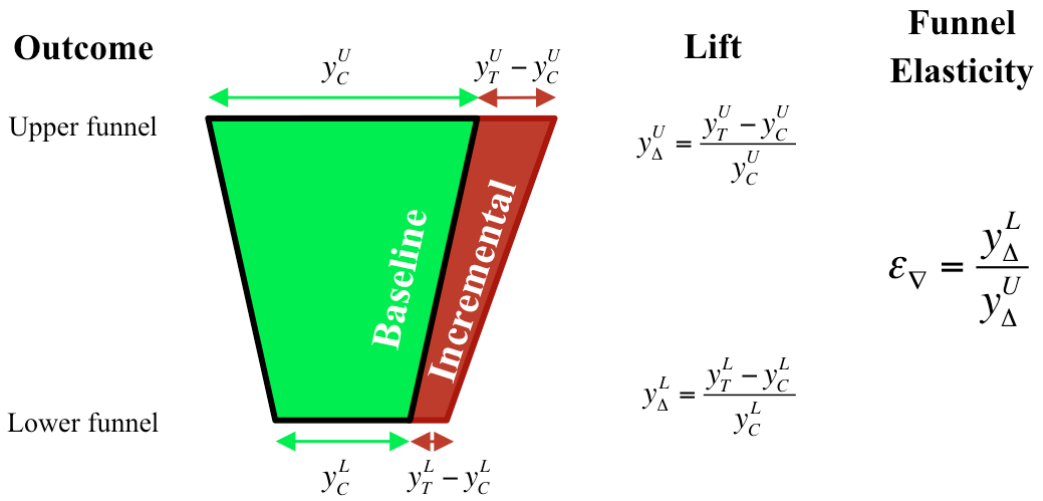
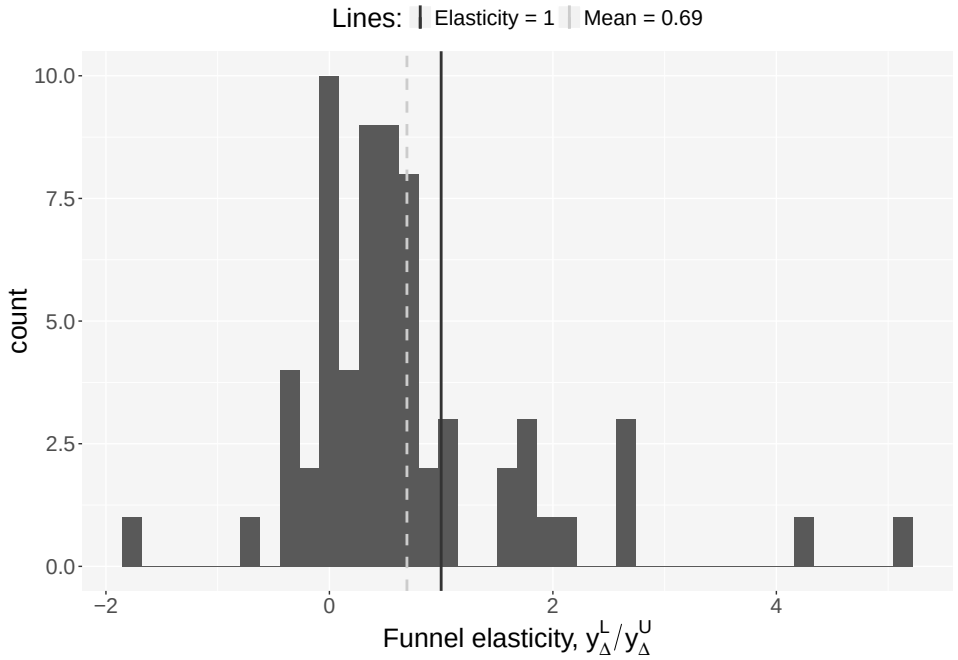
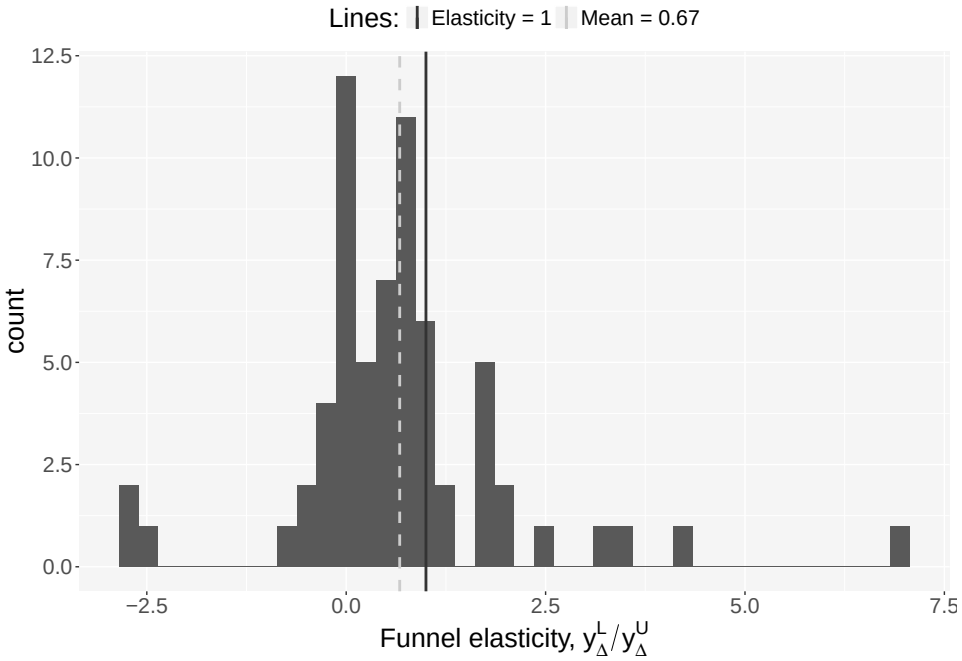


Figure 7: Extensive ad effectiveness funnel: Incremental visitors to converters



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

Figure 8: Ad effectiveness funnel: Incremental visits to conversions



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

Table 1: Summary of ad effectiveness large, multi-advertiser field experiments

Study	Medium	Experiment type	Outcome	Sample size (avg.)	# of studies	% significant†	Collective significance††
Blair (1987)	Television	Weight	Sales	?	20	30% (p=.05)	3.29 * 10^-4
Lodish et al. (1995a)	Television	Weight	Sales	~18,000	292	39.4% (p=.2)	2.35*10^-14
		Copy			96	27.1% (p=.2)	.057
MacInnis et al. (2002)	Television	Weight	Sales	?	47	53.2% (p=.05)	7.72* 10^-28
Hu et al. (2007)	Television	Holdout, Weight	Sales	<12,000* ;~32 markets	210	30% (p=.015)	4.31*10^-62
Hu et al. (2009)	Television	Holdout, Weight	Sales	<12,000*; ? markets	38	50% (p=.025)	8.25*10^-21
Kalyanam et al. (2015)	Search	Holdout	Sales	29 markets	14	71.4% (p=.025)	8.71*10^-14
Dai & Luca (2016)	Search	Holdout	Visits, Conversions	-	18,000	-	-
Sahni (2015, 2016)	Search-Display	Holdout, Weight	Visits	19,194	11	-	-
			Conversions			-	-
Goldfarb & Tucker (2011a)	Online display	Holdout	Survey	852	2,892	-	-
Bart et al. (2014)	Online display (mobile)	Holdout	Intention	740	54	29.6% (p=.025)	1.99*10^-13
			Attitude			20.4% (p=.025)	8.45*10^-08
Gordon et al. (2017)	Online display	Holdout	Conversions	36,431,611	15	66.7% (p=.025)	2.55*10^-13
Present study	Online display	Holdout	Visits	3,139,227	347	56.2% (p=.025)	7.38 * 10^-213
			Conversions	5,828,395	184	28.8% (p=.025)	2.93 * 10^-40

Notes: † $H_0 : lift \leq 0$; one-sided test at threshold p. †† Collective significance evaluated using binomial test with null hypothesis of i successes of N trials with Bernoulli probability p.

Table 2: Overall lift across 432 predicted ghost ad experiments

	Visits	Visitors	Conversions	Converters
Median lift estimate	16.6%	21.8%	8.1%	7.9%
[.10, .90]-quantile	[-1.1%, 213.6%]	[-0.2%, 284.4%]	[-8.9%, 83.4%]	[-7.2%, 67.5%]
[.025, .975]-quantile	[-8.1%, 780.9%]	[-4.8%, 830.3%]	[-29.5%, 258.8%]	[-26.2%, 252%]
Average lift estimate	1189.2%	699.7%	19.9%	19.8%
Standard error	(1098.9%)	(596.1%)	(10.8%)	(8.5%)
Weighted average lift estimate[††]	836.4%	534.9%	25.3%	34.4%
Standard error	(705.7%)	(384.1%)	(14.6%)	(20%)
Individual significance tests				
Reject 5% two-sided	202	249	57	65
Reject 2.5% one-sided (Lift <0)	7	8	4	5
Reject 2.5% one-sided (Lift >0)	195	241	53	60
Collective significance[†††]	$7.38 * 10^{-213}$	$1.43 * 10^{-296}$	$2.93 * 10^{-40}$	$6.11 * 10^{-49}$
N. of Experiments	347	347	184	184

Notes: [††]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 3: 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 4: 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.

Table 5: 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 6: 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.

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., 2017). 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. (2016) 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.