When Less is More: Data and Power in Advertising Experiments

Garrett A. Johnson, Randall A. Lewis, and David H. Reiley *

December 29, 2015

Abstract

Yahoo! Research partnered with a nationwide retailer to study the effects of online display advertising on both online and in-store purchases. We use a randomized field experiment on 3 million Yahoo! users who are also past customers of the retailer. We find statistically significant evidence that the retailer ads increase sales 3.6% relative to the control group. We show that control ads boost measurement precision by identifying and removing the half of in-campaign sales data that is unaffected by the ads. Less data gives us 31% more precision in our estimates—equivalent to increasing our sample to 5.3 million users. By contrast, we only improve precision by 5% when we include additional covariate data to reduce the residual variance in our experimental regression. The covariate-adjustment strategy disappoints despite exceptional consumer-level data including demographics, ad exposure levels, and two years' worth of past purchase history.

Keywords: advertising effectiveness, field experiments, digital advertising

*Johnson: Simon Business School, University of Rochester, <Garrett.Johnson@Simon.Rochester.edu>. Lewis: <randall@econinformatics.com>, Netflix. Reiley: <david@davidreiley.com>, Pandora Media, Inc. and University of California at Berkeley. The authors performed this research while employed at Yahoo! Research. We acknowledge Yahoo! both for its financial support and for its support of unbiased scientific research through its ex-ante decision to allow us to publish the results of the experiment no matter how they turned out. We would like to thank Valter Sciarillo, Jennifer Grosser Carroll, Taylor Schreiner, Eddie Babcock, Iwan Sakran, and many others at Yahoo! and at our client retailer who made this research possible. We thank Eric Anderson, Susan Athey, Mitch Lovett, Ben Jones, Preston McAfee, Rob Porter, Justin Rao, Elie Tamer, and Florian Zettelmeyer for helpful discussions and Mary Kraus for proofreading.

1 Introduction

From 2007 to 2011, Yahoo! Research conducted a number of large-scale controlled experiments to measure the effects of online display advertising on retail sales. The experiment reported in this paper represents the best of these, incorporating wisdom accumulated during other experiments and breaking new ground on scale. We examine two consecutive week-long ad campaigns that target three million of an apparel retailer's existing customers, matching subsequent in-store and online retail purchases with advertising exposure at the individual level. These retail image ads show attractive photos of featured items, but no prices, detailed product information, or calls to action. Our experimental estimates suggest the retailer ads increased sales by 3.6% and that the campaigns were profitable.

The Yahoo! Research experiments demonstrate that even large ad experiments have low statistical power. Lewis and Rao (2015) critically examine these Yahoo! Research advertising experiments. They point out a 'signal-to-noise' problem: the plausible mean impact of advertising is typically a small fraction of the variance in sales across consumers for occasionally purchased products. Such ad experiments have low statistical power and may require millions of observations—across consumers, campaigns, or time—to detect a profitable impact. With this bleak reality in mind, we saw two opportunities to reduce the variance in the outcome variable and thus alleviate the power problem. First, we add rich covariates to our dataset, including a two-year history of purchases prior to the experiment, in order to reduce residual variance and, hence, standard errors. Second, we improve upon the precision of intent-to-treat estimates by using control ads in the control group in order to exclude the noise from purchases by users who are not exposed to an ad.

In this paper, we employ both approaches by adding covariates to our estimator of the treatment effect and subtracting irrelevant observations via the use of control ads. We find that the control-ad approach is more effective: it improves the precision of our ad lift estimates by 31%, whereas the covariates approach only improves precision by 5%. When combined, the two approaches produce a 34% improvement in precision, equivalent to increasing our sample size by 80% to 5.6 million users. Conditioning on covariates underwhelms despite over 200 user-level covariates,

including ad exposure intensity, user demographics, retailer customer categories, and two years of past sales history. Covariates struggle to predict sales in the campaign because purchases at the retailer are occasional, unpredictable and highly variable in amount conditional on purchase. The control-ad approach requires that the ad platform delivers the retailer ads in the treatment group symmetrically to the control ads in the control group, which we verify in the experiment. One quarter of the control-ad improvement results from a related methodological innovation: we discard sales that occur in the campaign prior to a user's first impression.

Retailers account for a large share of online display ads (8% of total impressions, according to comScore 2013), but multi-channel retailers face challenges in measuring the in-store sales impact of their online advertising. To help solve this problem, Yahoo! Research partnered with five retailers and with third-party data-matching services in order to link consumer data on Yahoo! ad views with a retailer's in-store and offline sales data. This proved to be a nontrivial task: a number of the experiments encountered technical problems, rendering data unusable or statistically underpowered. Such database-match campaigns allow advertisers to target past customers; this capability is available on major ad platforms like Facebook and Twitter as well as Yahoo!.

Among the Yahoo! Research experiments attempted, the experiment reported in this paper stands out for its large size and spending as well as its superior experimental design and execution. Three design improvements distinguish this experiment from our previous best experiment with the same retailer, described in Lewis and Reiley (2014). First, our experiment includes 3 million eligible users—double the size of its predecessor—and balances the treatment/control group split. Second, our experiment includes control ads, which boost the precision of our estimates. Third, the experiment features exceptional data on consumers, which also boost precision. With these advantages, our experiment delivers significant evidence that the ad campaigns increased sales. In contrast, Lewis and Reiley (2014) do not obtain statistically significant experimental estimates without making a difference-in-differences assumption that is difficult to precisely test.

We designed our experiment to examine the impact of ad exposure frequency. The experiment includes a 'Full' treatment group that is exposed to the retailer's ads and a 'Control' group that is

exposed to unrelated control ads. We also included a 'Half' treatment group that sees a retailer or control ad with equal probability at each ad viewing opportunity. Our experimental estimates suggest an average sales lift of $\$0.477$ on the Full treatment group and $\$0.221$ on the Half treatment group, which represent a 3.6% and a 1.7% increase over the Control group's sales.

The rest of this paper is organized as follows. The next section reviews the experimental ad effectiveness literature. Section 3 describes our experimental design, while the fourth section provides descriptive statistics and validates the experiment. The fifth section presents our measurements of the causal effects of the advertising. The final section concludes.

2 Literature Review

Yahoo! Research emphasized the use of controlled experiments in order to avoid endogeneity between advertising and consumer behavior, having documented serious bias from methods commonly used in the advertising industry to analyze observational data. Naive observational estimates that compare endogenously exposed versus unexposed users may overstate ad effectiveness by orders of magnitude (Lewis, Rao and Reiley, 2011), largely not overlap (Hoban and Bucklin, 2015), or even have the wrong sign relative to experimental estimates (Lewis and Reiley, 2014). Advertisers' advertising choices can induce bias, for instance, by targeting customers who are more likely to purchase or by targeting times like Christmas when customers purchase more. Lewis, Rao and Reiley (2011) show that online consumer choices can also induce bias (what they term 'activity bias') because consumers' activity is correlated across websites without that correlation necessarily being causal.

Controlled experiments remain rare in the advertising industry. This dearth of experiments seems surprising given the dollars at stake: advertising represents between 1% and 2% of global GDP (Bughin and Spittaels, 2012), and U.S. online display advertising revenues alone reached $7.9 billion in 2013, excluding mobile (IAB, 2014). We believe that the potential pitfalls of endogeneity and selection bias in observational studies are not well appreciated by industry analysts, despite

a report commissioned by the Interactive Advertising Bureau (Lavrakas, 2010) affirming experiments as the gold standard for measurement. To be sure, observational studies can provide value by exploiting natural sources of exogenous variation in advertising (see e.g. Hartmann and Klapper 2014; Shapiro 2015; Sinkinson and Starc 2015; Stephens-Davidowitz et al. 2015). The cited studies rely on a combination of high ad spend and a type of natural experiment specific to television advertising. Most of them also rely on aggregation of multiple campaigns to get statistically significant results, combining many experiments instead of measuring the effects of a single campaign as in this paper. We find experiments particularly valuable in the digital setting, where randomization can be conducted at the level of the individual consumer.

Lewis and Rao (2015) describe the severe statistical power problem in ad experiments that past studies resolve in different ways. One solution is to examine the effect of ads on less sparse or noisy outcome variables than sales. For instance, multiple online ad studies use survey outcomes like purchase intent and brand recall (e.g. Goldfarb and Tucker 2011; Bart et al. 2014). Other online ad studies use indicator outcomes like clicks (Lewis 2010; Bakshy et al. 2012), site visits, sales leads (Sahni 2015a), transactions (Lambrecht and Tucker 2013) or some combination of these purchase funnel indicators (Hoban and Bucklin 2015). We are interested in measuring the effect on online and in-store sales, thoug this is a harder estimation problem, because this allows us to evaluate the short-run return on investment for the campaign. Another solution to the power problem is to study settings with large ad effects like in online search where users are often seeking out competing advertisers with an intent to purchase (Sahni, 2015a; Kalyanam et al., 2015), though not if the users already have an advertiser in mind (Blake et al., 2015). Instead, we are interested in the effects of image ads on users who receive ads while browsing unrelated content; our efforts to increase power are therefore crucial.

Studies that examine purchase outcomes are rare and often gain power by combining studies. Lodish et al. (1995) pioneered split-cable television experiments with a panel of 3,000 households with advertising treatment matched to household-level purchases of consumer packaged goods. With merely thousands of consumers per experiment, these 600 experiments individually lack sta-

tistical power—the majority were statistically insignificant—but collectively demonstrate a significant ad effect in a meta-study (Lodish et al., 1995; Hu et al., 2007). Kalyanam et al. (2015) demonstrate that product-category search advertising increases offline retail sales using a meta-study of 15 experiments and 13 retailers. Kalyanam et al. (2015) vary advertising at the DMA level then compare sales at treated stores to similar counterparts. Again, these studies are collectively significant though only half are individually significant. Simester et al. (2009) is an exception—their single experiment finds statistically significant results in the catalog setting, where customer-level advertising is readily linked to customer-level sales.

Control ads have previously been used in ad experiments to improve measurement precision or to identify a set of holdout users. Control ads take the form of charity ads (see e.g. Yildiz and Narayanan 2013; Hoban and Bucklin 2015), ads from unrelated companies (see e.g. Lewis 2010; Lewis and Nguyen 2015), house ads promoting the platform (see e.g. present study and Sahni 2015a), or blank ads (see e.g. Goldfarb and Tucker 2011; Bart et al. 2014).

Though ours is not the first to employ control ads, the present study stands out for its rich individual-level covariates including user demographics, ad exposure, retailer RFM categories, and two years of historical sales separately by online and offline channel. Online ad experiments typically only have covariates for ad exposure and browsing behaviour during the experiment or a few weeks before (Hoban and Bucklin, 2015; Lambrecht and Tucker, 2013; Sahni, 2015b) or self-reported survey covariates (Goldfarb and Tucker, 2011; Bart et al., 2014). Other online ad experiments do not include past behavior or demographics as covariates—preferring to present simple experimental difference results (Lewis and Reiley, 2014; Sahni, 2015a). Many market-level ad experiments use past market-level sales or demographic covariates not to improve precision but to try to reduce bias in comparing the treated sales to control markets or predicted sales (Eastlack Jr and Rao, 1989; Lodish et al., 1995; Hu et al., 2007; Kalyanam et al., 2015). Simester et al. (2009) stand out for including RFM variables from fifteen years of sales history as well as the proximity to stores as covariates in their catalog experiment.

3 Experimental Design

The experiment measures Yahoo! advertising effectiveness for a national apparel retailer. The experiment took place during two consecutive weeks in spring 2010. In each week, the advertisements featured a different line of branded clothing. The experimental subjects were randomly assigned to treatment groups that remained constant for both weeks. A confidentiality agreement prevents us from naming either the retailer or the featured product lines.

To investigate the effects of exposure frequency, our experiment uses three treatment groups that vary the treatment intensity. The 'Full' treatment group is exposed to the retailer's ads while the 'Control' group is exposed to unrelated control ads. A third, 'Half' treatment group is, on average, exposed to half of the retailer and control ads of the other two groups. We implement this design by booking 20 million retailer ads for the Full group, 20 million control ads for the Control group, and both 10 million retailer ads and 10 million control ads for the Half group.[1] Each experimental ad exposure in the Half group therefore has a 50% chance of being a retailer ad and a 50% chance of being a control ad. This experimental design enables us to investigate the impact of doubling the number of impressions in a campaign. Doubling the size of the campaign increases ad delivery on two margins: 1) showing more ads to the same consumers (the intensive margin) and 2) increasing the number of consumers reached, here by 8% (the extensive margin). The average ad frequency in the Half group is comparable to a typical campaign for this retailer on Yahoo!.

The retailer ads are primarily image advertising and are co-branded with apparel firms. The ads display the store brand, the brand of the featured product line, and photographs of the advertised clothing line worn by attractive models. The ads do not include any price or promotion information. Figure 1 presents an illustrative example with the retailer Target and the designer Missoni—neither of which is involved in the experiment. The creative content of each ad impression is dynamic, with slideshow-style transitions between still photographs and text. Campaign 1 includes three

[1] These campaigns were purchased on an impression, not click, basis. This avoids distortions induced by ad servers optimizing delivery patterns in the treatment and control groups due to the retailer and control ads having different click-through rates. Such distortions can render the control ads unsuitable for tagging counterfactual exposures in the control group. See Johnson et al. (2015) for a discussion and an control-ad alternative that is robust to an ad server's campaign optimization.

different ads in an equal-probability rotation: men's apparel, women's apparel, and women's shoes. Campaign 2 advertises a product line from a different manufacturer and features women's apparel. The control ads advertise Yahoo! Search and are depicted in Figure 2.

The experiment implements a 'database-match' campaign in which the subjects are existing customers of both Yahoo! and the retailer. Prior to the experiment, a third-party data-collection firm matched the customer databases of Yahoo! and the retailer using the customer's name and either terrestrial or email address. Leveraging additional customer records, the third party doubled the number of matched customers from the 1.6 million customers studied by Lewis and Reiley (2014) to 3.1 million in the present experiment. After the experiment ended, the third-party firm combined the retailer's sales data and the Yahoo! advertising data and removed identifying information to protect customer privacy. The retailer believes that its sales data correctly records the individual customer for more than 90% of all purchases. Matching customers sales and ad exposure data frequently results in a multiple-identifiers matching problem. Our original data source also suffered from such a problem: the data contain more retailer than Yahoo! identifiers. For simplicity, we focus our analysis on the 3.1 million users who were uniquely matched (see Appendix A.1 for details). The experiment thus measures the causal effects of advertising on this particular intersection of Yahoo!'s users and the retailer's customers.

Database-match campaigns like the one in our experiment are used by advertisers to target existing customers or to prospect for new customers. This service is available on major platforms like Yahoo!, Twitter (Lynn, 2014), Google (Ramaswamy, 2015), and Facebook, which even allows targeting for buyers of automobiles and consumer packaged goods (Datalogix, 2014). Though the experiment's campaign targets existing customers, only 41% of users in the sample have transacted with the retailer in the past eight weeks, and 3.6% have not transacted in the previous two years. Thus, database-match campaigns reach both active and inactive customers whereas retargeting campaigns (see e.g. Lambrecht and Tucker 2013) reach only consumers who visited the retailer's website recently. Unlike retargeting, database-match campaigns can target users based on their offline sales and less recent sales. Moreover, database-match campaigns target logged-in

users who are linked to their offline identities. The fact that we deliver ads only to logged-in users represents a great advantage relative to most online-advertising measurement efforts, which suffer from problems like impersistent cookie identifiers and the difficulty of linking user activity across multiple browsers and devices. Users in the campaign receive ads as they browse mostly unrelated content on Yahoo!. As such, the ads are much closer to traditional print and television ads than to online search ads, since users express intent through their queries in the latter case.

Many users in the experiment *do not see an ad* because they either do not visit Yahoo! at all or do not browse enough pages on Yahoo! during the campaigns. The experiment employs control ads to identify the counterfactual treated users in the control group who would have received the retailer's ads. The control ads also tell us the number of counterfactual ad exposures that a consumer would see if they had been assigned to the Full group. The experiment's retailer and control ad campaigns join all other competing ad campaigns in Yahoo's ad server, which selects each ad on each pageview on Yahoo! subject to ad supply and demand constraints, campaign-targeting criteria, and revenue-optimization goals. The experiment's ads ran on about 8% of Yahoo! pageviews and appeared on all Yahoo! properties including Mail, News, and Autos. The ads have four rectangular formats. The campaigns were identically configured for all three treatment groups, except for the ad creatives featuring Yahoo! versus the retailer.

To demonstrate the limits of our experiment, we calculate that our experiment has the statistical power to reject the null hypothesis that advertising has no impact on sales 79% of the time in the Full group and 34% in the Half group. In this calculation, we consider the alternative hypothesis that the advertiser receives a 50% return on its advertising investment. The alternative hypothesis implies an average treatment effect on the treated of $0.51 in the Full treatment group, given the $0.17 cost of display ads and assuming a 50% contribution margin for the retailer. The standard deviation of sales is $125 for the two-week campaign and the sample size is 570,000 in each of the Full and Control treatment groups. In Section 5, we present methods that reduce the standard deviation of sales to $111, without which our power would be much lower (49% instead of 79%). A comparable calculation for Lewis and Reiley (2014) reveals that its study had only 47% power. If

we seek to detect longer-run ad effects, this compounds the statistical power problem. In particular, if the hypothesized $0.51 lift occurs against the background noise of more than just the two weeks of sales during the campaign, our statistical power will be reduced.

4 Data & Experimental Validation

This section describes our data and demonstrates the validity of our experimental randomization. In particular, we demonstrate that the distribution of user characteristics, pre-treatment outcomes, and experimental ad exposures are equivalent across treatment groups in accordance with experimental best practices (see Gerber and Green (2012) Ch. 4). In other experiments not reported here, our randomization checks have often helped us to uncover execution errors such as incorrect matching of sales to advertising data, failure to create identical targeting between treatment and control ads, or unexpected selection bias generated by an advertising auction. Our randomization checks here validate that the treatment and control groups are equivalent but for the exposure to retailer advertising. Appendix A.1 details the source of our data and some key variables.

We verify the randomization by testing the distribution of user characteristics and pre-treatment outcomes. Table 1 provides summary statistics for the experiment. In Table 1's right-most column, F-tests of equality of means of the variables by treatment group are not rejected, which is consistent with a valid experimental randomization. Over three million customers were evenly assigned to one of our three treatment groups: Full, Half, or Control. In each treatment group 68.5% of customers are female, the average age is 43.6 years, and customers viewed an average of 245 web pages on Yahoo! during the campaign. Customers spent an average of $19.23 at the retailer during the two weeks prior to the experiment and $857.53 in the two years beforehand. Figure 3 shows that the distribution of average weekly sales over the two years prior to the experiment are essentially identical across all three treatment groups.

Ad distribution tests are critical in ad experiments with control ads to demonstrate that the total number of experimental ads (treatment + control) are delivered symmetrically across treatment

and control groups. Ad platforms are complex and may fail such tests for many reasons including the use of click- or action-optimized campaigns. In such cases, the ad platform will optimize ad delivery such that the retailer-ad exposed users look different from control-ad exposed users. Our experiment uses a simple reach-based campaign and avoids such problems. In Table 1, we see that the experiment delivers advertisements evenly across treatment groups. Ad exposure depends on a user's browsing activity during the two weeks of the campaign; 55% of users were exposed to an experimental ad. Figure 4 shows that the distribution of total ad views (both retailer and control) across the three treatments is identical and the F-test of equality reveals no significant differences in average impressions in Table 1. The distribution of ad views is highly skewed right, so that the mean is 33 while the median is 15 among exposed users. As expected, the Half treatment group sees an even split of retailer and control ads.

5 Results

In the results section, we show the experimental estimates for the sales lift during the two weeks of the ad campaign. We present methods that improve the statistical precision of our estimates in this low-powered setting. Our preferred experimental estimates suggest that consumers who were exposed to the retailer's ad saw their average sales increase by $\$0.477$ (3.6%) in the Full treatment group and $\$0.221$ (1.7%) in the Half treatment group. In Appendix A.2, we separate the effect of advertising by campaign, sales channel, and shopping trips. We find that the majority of the total treatment effect is attributable to the in-store rather than online sales channel and estimate a 1.8% increase in shopping trips in the Full group. Appendix A.2 also shows that including sales after the campaign increases the estimates of the overall causal effects, which allays the concern that the in-campaign estimates might merely reflect intertemporal substitution of purchases from future to present.

Table 2 presents regression estimates of the average effect of treatment on the treated (TOT) for the impact of advertising on consumer purchases during the two-week experiment. In particular,

we contrast two approaches for increasing the precision of the estimates without inducing bias: the control-ad and the covariates approaches. The control-ad approach works by pruning components of the outcome data that cannot be influenced by advertising. The covariates approach introduces covariates into the experimental linear regression to reduce the residual variance of the outcome variable. In all, we improve the precision of our estimates—or shrink the standard errors—by 34% on average.

As a baseline, we begin with the Intent-to-Treat (ITT) estimates from which we derive the indirect TOT estimate. The indirect TOT estimator takes the treatment-control difference for the entire sample of 3.1 million consumers (ITT estimate) and divides by the treatment probability (the 55.4% exposed subsample).[2] The indirect TOT estimator relies on the fact that outcomes among untreated subjects have an expected experimental difference of zero. However, variance in outcomes among untreated subjects adds noise (but no signal) to the estimator. As column (1) of Table 2 indicates, the indirect TOT estimator yields a \$0.67 average sales lift (s.e. \$0.32) in the Full treatment group and an average lift of \$0.03 (s.e. \$0.31) in the Half group. Whereas Lewis and Reiley (2014) estimate the TOT indirectly out of necessity, this experiment uses control ads to isolate the counterfactual treated sample in the Control group.

While the retailer ads identify the treated among the treatment group, we need control ads to identify the counterfactual treated users in the control group who would have seen the ad. Table 2's column (2) presents the direct TOT regression estimate on the treated sample. The treated sample are those users who see any of the experiment's retailer or control ads during the two weeks of the campaign. The direct TOT estimator increases precision by pruning the untreated subsample that contributes only noise to the estimator. By using the control ads to filter out the unexposed users, we improve the precision of the experimental estimates—or decrease the standard errors—by 25% on average.

We propose a second use for control ads that further improves precision: we omit purchases

[2]This is numerically equivalent to computing a local average treatment effect by using the random assignment as an instrument for treatment. The unscaled, intent-to-treat estimates are \$0.37 for the Full group and \$0.01 for the Half group.

that occur prior to a consumer's first experimental ad exposure, because ads cannot influence sales until the user receives the ad. Key to achieving this result was obtaining daily data on individual consumer purchases during the campaign. Because the control ads identify the counterfactual pre-treatment sales in the Control group, we can exclude purchases on days before the first ad impression is seen by a given consumer. This method requires that the ad platform delivers the experimental ads symmetrically across treatment groups, which we verified in Section 4. Table 2's column (3) uses both the control ads and daily sales data to further prune the data and boost precision by another 8%. In all, by trimming sales of unexposed users and pre-treatment sales, we remove 52.4% of total purchases during the campaign, which in turn shrinks our confidence intervals by 31%. As our title suggests, less data means more precise estimates.

Next, we apply our covariates approach to increase the precision of our estimates. Adding covariates to the TOT regression improves precision by reducing the unexplained variance in the dependent variable.[3] Specifically, these covariates include: demographics, retailer-defined customer segments, consumer sales history, and ad-exposure intensity. The demographic covariates include indicator variables for gender, year of age, and state of residence as well as a scalar variable for the time since the consumer signed up with Yahoo!. The retailer-defined RFM customer segments include Recency of last purchase, Frequency of past purchases, and Monetary value (total lifetime spending at the retailer). We include 54 variables capturing two years of individual-level past sales amount data and pre-treatment sales amount during the campaign, separately for online and offline sales. The ad-exposure-intensity covariates are a set of indicator variables for the total number of experimental ads delivered and for the day of the consumer's first ad exposure. To the extent that shopping behavior is correlated with current online browsing activity, the exposure intensity fixed effects will improve efficiency. In columns (4)-(7), we gradually add these different covariate types. The historical sales and customer category covariates account for nearly all of the 5% improvement in precision. The demographic and exposure-intensity covariates provide almost no precision improvement. In total, the full regression model in column (7) includes 236 covariates

[3] A regression model with a given R^2 reduces the standard errors of our treatment effect estimates by $\approx 1 - \sqrt{1 - R^2}$.

in addition to the Full and Half treatment dummy variables but only achieves a R^2 of 0.09.

Note that these results apply to a retailer of occasionally purchased goods; past purchase covariates may prove more useful for frequently purchased goods. Apparel retailer sales are difficult to predict because—unlike CPG goods like milk—transactions are infrequent and highly variable conditional on purchase. 7.7% of control group users transacted during the campaign after their first ad exposure, and the standard deviation in sales conditional on transacting is $385. Table 5 column (2) reveals that even the transaction indicator outcome is difficult to predict, as the R^2 there is only 0.15. In an unpublished study with another apparel retailer, we find comparable R^2 for sales regressions that include past sales covariates: the R^2 varies between 0.017 and 0.073 across seven campaigns.

Including covariates improves precision by 5% (columns 3–7) across Full and Half groups, while pruning irrelevant data improves precision by 31% (columns 1–3) on average. Put another way, the experiment would have to be 10% larger without covariates or 71% larger without control ads to maintain the same measurement precision. Collectively, covariates and control ads boost precision by 34%, which would otherwise require an 80% larger sample of 5.6 million users as well as 80% larger ad spend. Though we find covariates to be less effective at improving precision, covariates may be inexpensive to include and serve to validate the experimental randomization. Control ads are expensive, since one has to pay for additional advertising inventory (Johnson et al. 2015 suggest a low-cost alternative), but they facilitate data pruning and thereby improve precision five times more than covariates. This is an example where less is more: subtracting data proves more valuable than adding data.

Our preferred experimental estimator—in column (7) of Table 2—measures a $\$0.477$ (s.e.: $\$0.204$) increase in average sales from the ads in the Full group and a $\$0.221$ (s.e. $\$0.209$) increase in the Half group. Though all the estimates in Table 2 are unbiased, we prefer the estimates in column (7) as they are the most precise. These point estimates with covariates are more conservative, as they are smaller than those in column (3). The point estimates in column (7) likely fall because they account for differences like the slightly lower sales in the Control group in the two weeks before

treatment (see Table 1). The preferred Full treatment effect is statistically significant at the 5% level (p-value: 0.020) and represents a 3.6% sales lift over the Control group. The Half treatment effect is not significantly different from zero (p-value: 0.289), and the joint F-test is marginally significant (p-value: 0.065). The point estimates indicate that doubling the advertising approximately doubles the effect on purchases, but the precision of these estimates is too low for us to have great confidence in the result. Recall from our power calculations in Section 3 that this is the most likely outcome: under our proposed alternative hypothesis of a 50

Our estimates imply a relatively high elasticity of sales with respect to advertising of about 0.19. Elasticity is the the derivative of sales with respect to advertising, multiplied by the advertising-to-sales ratio. For the derivative of revenue with respect to advertising, we divide the incremental effect of ads in the Full group by the cost of the advertising ($4.60 CPM, with 33.4 ads delivered per person): (dR/dA)=$0.477/(33.4*$0.0046)=3.10. Elasticity depends crucially on the interpretation of the advertising-to-sales ratio, which we interpret as the retailer's total advertising across channels. Since we were not given the advertising-to-sales ratio of the retailer, we impute this ratio to be 6% from the financial filings of a competitor. Our short-run elasticity of 3.1*6%=0.19 exceeds the average elasticity of 0.12 and median of 0.05 given in the meta-study by Sethuraman et al. (2011), though we note that they exclude all experimental estimates. The TV ad experiment meta-study by Hu et al. (2007) calculates the elasticity of online display ads to be 0.11 using another approach: they use the campaign's ad weight over the during-campaign sales as the ads-to-sales ratio. By this method, our elasticity estimate is much lower at about 0.017.[4]

To make decisions about advertising, managers want not only to establish a revenue lift, but also calculate return on investment. The profitability of advertising depends not only on the elasticity of advertising, but also on the retailer's gross profit margin and the cost of the ads. Given 570,000 exposed users in each of the three treatment groups, our point estimates indicate that the Full

[4]We calculate the 'local' elasticity measure at the Half group ad intensity level since this level of intensity is normal for the retailer. We use the increase in incremental sales from the Half to the Full group as our derivative ($.477 - $.221) / (33.42*$0.0046 - 16.69*$0.0046), using this ad campaign's cost of $4.60 per thousand impressions (CPM). Since sales in a Half group among those exposed to 17 ads ($\pm$2 ads) is $15.20, we use $15.20/(16.7*$0.0046) as the 'local' ads-to-sales ratio.

campaign increased retail purchases by a dollar value of $\$273,000 \pm 229,000$, while the Half campaign increased purchases by $\$126,000 \pm 234,000$, using 95% confidence intervals. Compared with costs of about $88,000 for the Full campaign and $44,000 for the Half campaign, these point estimates indicate incremental revenues of around three times the cost. We assume a contribution margin of 50% for the retailer's sales.[5] Our point estimates indicate a rate of return of 51% on the advertising dollars spent, but with a 95% confidence interval of [-101%, 204%].

Our short-run sales effect and corresponding rate-of-return estimates are likely conservative, as several factors attenuate the measurement. These factors include: 1) incomplete attribution of sales to a given consumer, 2) mismatching of consumer retail accounts to Yahoo! accounts, 3) logged-in exposures viewed by other household members, and 4) observing purchases for a time period that fails to cover all long-run effects of the advertising. Though short-run effects could outpace the long-run effects due to intertemporal substitution as in Simester et al. (2009), our estimates in Appendix A.2.1 that include sales for the two weeks post-treatment suggest a positive long-run effect.

6 Conclusion

This paper provides a helpful case study of an ad experiment for both academics and practitioners, one that may be useful in the classroom. We examine the managerial question: how do I measure the total effect of my online display advertising if most of my sales are offline? A field experiment yields an elegant solution to this problem when combined with CRM sales data. With the results from our large experiment, the retailer has confidence that their ads increased online and in-store sales. Despite the significant sales results and high ad elasticity measure, the confidence region for the profitability of the campaign includes zero. However, the point estimates indicate that the campaign was likely profitable in the short run, and we believe our estimates understate the true sales lift due to several measurement issues that attenuate the result and because our post-campaign

[5] We base this on conversations with retail experts, supplemented with references to similar retailers' financial statements.

estimates indicate that the campaign's benefits might carry over past the end of the measurement period.

This experiment highlights three nuances of the online display ad field experiments. First, this study highlights the limits to what can be learned from large ad experiments and thus the challenge of optimizing advertising. As Lewis and Rao (2015) elaborate, hundreds of millions of subjects may be required to experimentally evaluate a hypothesis positing a 10% profit on ad spending. By their nature, online display ad effect estimates are imprecise and managers should take point estimates with a grain of salt so as not to overreact to either good or bad news.

Second, covariate data serve two important roles in ad field experiments. Covariates improve the precision problem by decreasing the residual variance of the outcome variable. While covariates improved precision here by only 5%, this approach may perform better in settings with frequently purchased goods or in other ad media. In addition, covariates allow the experimentalist to validate the randomization as we do in Section 4, a best practice that is especially important in complex online ad systems.

Third, control ads increase precision — in our case, considerably more than do covariates. By trimming both unexposed users and outcomes prior to exposure, control ads increase the precision of our estimates by 31%—equivalent to increasing the number of subject by 71%. Nonetheless, control ads are seldom used because they are expensive and incompatible with cost-per-click (CPC) and cost-per-action (CPA) optimized campaigns. We note that the ghost ad methodology developed by Johnson et al. (2015) resolves these problems and still delivers the precision gains of control ads.

Unbiased field experiments measuring the effectiveness of advertising can contribute both to the science of consumer choice and to the advertising decisions of managers. We look forward to future research including studies that leverage either ghost-ad and intent-to-treat (see e.g. Gordon et al. 2015) experimentation platforms and illuminates how advertising influences consumer choice.

References

Bakshy, Eytan, Dean Eckles, Rong Yan, and Itamar Rosenn, "Social influence in social advertising: evidence from field experiments," in "Proceedings of the 13th ACM Conference on Electronic Commerce" 2012, pp. 146–161.

Bart, Yakov, Andrew T Stephen, and Miklos Sarvary, "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*, 2014.

Blake, Thomas, Chris Nosko, and Steven Tadelis, "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment," *Econometrica*, 2015, *83* (1), 155–174.

Bughin, Jacques and Steven Spittaels, "Advertising as an economic-growth engine," Technical Report, McKinsey & Company March 2012.

comScore, "2013 U.S. Digital Future in Focus," Technical Report, comScore 2013.

Datalogix, "Datalogix Announces Facebook Partner Categories for CPG, Retail and Automotive Brands," April 2014.

Gerber, Alan S and Donald P Green, *Field experiments: Design, analysis, and interpretation*, WW Norton, 2012.

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

Gordon, Brett R., Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky, "A Comparison Of Experimental And Observational Approaches To Advertising Measurement: Evidence from Big Field Experiments at Facebook," 2015.

Hartmann, Wesley R and Daniel Klapper, "Do superbowl ads affect brand share," Technical Report, SSRN Working Paper 2014.

Hoban, Paul R. and Randolph E. Bucklin, "Effects of Internet Display Advertising in the Purchase Funnel: Model-Based Insights from a Randomized Field Experiment," *Journal of Marketing Research*, 2015, *52* (3), 375–393.

Hu, Ye, Leonard M Lodish, and Abba M Krieger, "An analysis of real world TV advertising tests: A 15-year update," *Journal of Advertising Research*, 2007, *47* (3), 341.

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

Johnson, Garrett A, Randall A Lewis, and Elmar I Nubbemeyer, "Ghost Ads: Improving the Economics of Measuring Ad Effectiveness," *Available at SSRN*, 2015.

Jr, J.O. Eastlack and A.G. Rao, "Advertising experiments at the Campbell Soup company," *Marketing Science*, 1989, pp. 57–71.

Kalyanam, Kirthi, John McAteer, Jonathan Marek, James Hodges, and Lifeng Lin, "Cross Channel Effects of Search Engine Advertising on Brick and Mortar Retail Sales: Insights from Multiple Large Scale Field Experiments on Google.com," 2015.

Lambrecht, Anja and Catherine Tucker, "When does retargeting work? Information specificity in online advertising," *Journal of Marketing Research*, 2013, *50* (5), 561–576.

Lavrakas, Paul J, "An evaluation of methods used to assess the effectiveness of advertising on the internet," *Interactive Advertising Bureau Research Papers*, 2010.

Lewis, Randall A., "Where's the "Wear-Out?": Online Display Ads and the Impact of Frequency." PhD dissertation, MIT Dept of Economics 2010.

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

_ **and Justin M. Rao**, "The Unfavorable Economics of Measuring the Returns to Advertising," *Quarterly Journal of Economics (forthcoming)*, 2015.

_ **,** _ **, and David H. Reiley**, "Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising," in "Proceedings of the 20th international conference on World wide web" ACM 2011, pp. 157–166.

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

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

Lynn, Kelton, "Twitter Blog: New ways to create and use tailored audiences," 2014.

Ramaswamy, Sridhar, "Google AdWords Blog: Google brings you closer to your customers in the moments that matter," 2015.

Sahni, Navdeep, "Advertising Spillovers: Field Experimental Evidence and Implications for Returns from Advertising," January 2015.

_ , "Effect of temporal spacing between advertising exposures: evidence from online field experiments," 2015.

Sethuraman, Raj, Gerard J Tellis, and Richard A Briesch, "How Well Does Advertising Work? Generalizations from Meta-Analysis of Brand Advertising Elasticities.," *Journal of Marketing Research*, 2011, *48* (3), 457 – 471.

Shapiro, Bradley, "Positive spillovers and free riding in advertising of prescription pharmaceuticals: The case of antidepressants," *Available at SSRN 2477877*, 2015.

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

Sinkinson, Michael and Amanda Starc, "Ask Your Doctor? Direct-to-Consumer Advertising of Pharmaceuticals," 2015.

Stephens-Davidowitz, Seth, Hal Varian, and Michael D Smith, "Super Returns to Super Bowl Ads?," 2015.

Yildiz, T. and S. Narayanan, "Star digital: Assessing the effectiveness of display advertising," *Harvard Business Review: Case Study*, March 2013.

7 Figures & Tables

Table 1: Summary Statistics & Experimental Validation

	Treatment Group			
	Full	Half	Control	*p*-value
Sample size	1,032,204	1,032,074	1,032,299	0.988
Female (mean)	68.5%	68.5%	68.5%	0.794
Age (mean)	43.6	43.6	43.6	0.607
Yahoo! page views[a] (mean)	245.8	244.4	243.5	0.132[d]
Pre-Treatment sales (2 years, mean)	$857.74	$859.30	$855.54	0.475
Pre-Treatment sales (2 weeks, mean)	$19.34	$19.24	$19.10	0.517
Treated Subsample				
Exposed sample	572,574	571,222	570,908	0.254
Yahoo! page views (mean)	412.2	411.5	410.1	0.108
Ad views (mean)	33.42	33.41	33.66	0.164
Ad views (median)	15	15	15	
Retailer ad views (mean)	33.42	16.69	-	0.801
Control ad views (mean)	-	16.72	33.66	0.165
Retailer ad click-through rate[b]	0.19%	0.24%	-	
Retailer ad clicker rate[c]	4.91%	3.39%	-	

Notes: Sample includes only those customers who are uniquely matched to a single Yahoo! user identifier. [a]Webpage views on Yahoo! properties during the two weeks of the campaign. [b]The click-through rate is the quotient of total ad clicks and views. [c]The clicker rate is the proportion of users exposed to the ad who click on it. [d]Here we include pageviews in the 2 weeks prior to the experiment in the regression to reduce the impact of outliers.

Table 2: Effect of Advertising on Sales: Refinements in Precision

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Subset of Users[a]	Everyone	Treated	Treated	Treated	Treated	Treated	Treated
Sales After First Ad Exposure[b]			x	x	x	x	x
Full Treatment ($)	0.673**	0.525**	0.559**	0.553**	0.535**	0.486**	0.477**
	(0.317)	(0.237)	(0.217)	(0.217)	(0.213)	(0.204)	(0.204)
Half Treatment ($)	0.0248	0.189	0.307	0.307	0.239	0.241	0.221
	(0.311)	(0.235)	(0.217)	(0.217)	(0.212)	(0.209)	(0.209)
Constant ($)	15.52***	15.53***	13.17***				
	(0.122)	(0.166)	(0.154)				
Covariates							
Demographics (126 variables)[c]				x	x	x	x
Customer categories (13 variables)[d]					x	x	x
Past sales, 2 years (54 variables)[e]						x	x
Exposure intensity (43 variables)[f]							x
Observations	3,096,577	1,714,704	1,714,704	1,714,704	1,714,704	1,714,704	1,714,704
R^2	0.000	0.000	0.000	0.001	0.042	0.090	0.091

Average effect of Treatment on the Treated estimates. Dependent variable is sales during the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Demographic covariates include individual gender, age, state dummies as well as the user's tenure as a Yahoo! customer. [d]The retailer customer category covariates include categorical variables for recency of last purchase, customer loyalty, and lifetime customer spending. [e]Two-year sales of pre-treatment—both online and in-store sales amounts—at the weekly level except for aggregate sales for weeks 9–44 and 61–104 (54 variables in total). For models that use sales after the first ad exposure as the outcome variable, we include sales from the beginning of the campaign to that first exposure. [f]The exposure intensity covariates include fixed effects for the day of the first ad exposure and the number of total exposures (retailer or control) for 1 to 30 separately and a single indicator for >30.

Figure 1: Co-Branded Retailer-Designer Example (Experiment Uses neither Target nor Missoni)

Figure 2: Control Ad: Promoting Yahoo! Search

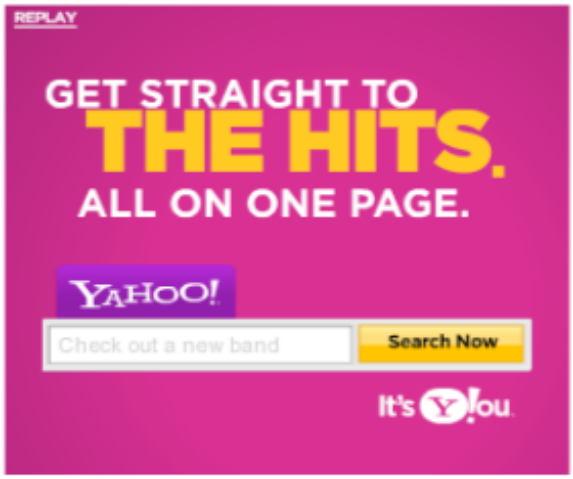


Figure 3: Histogram of Average Weekly Purchases During the Two Years Before the Experiment

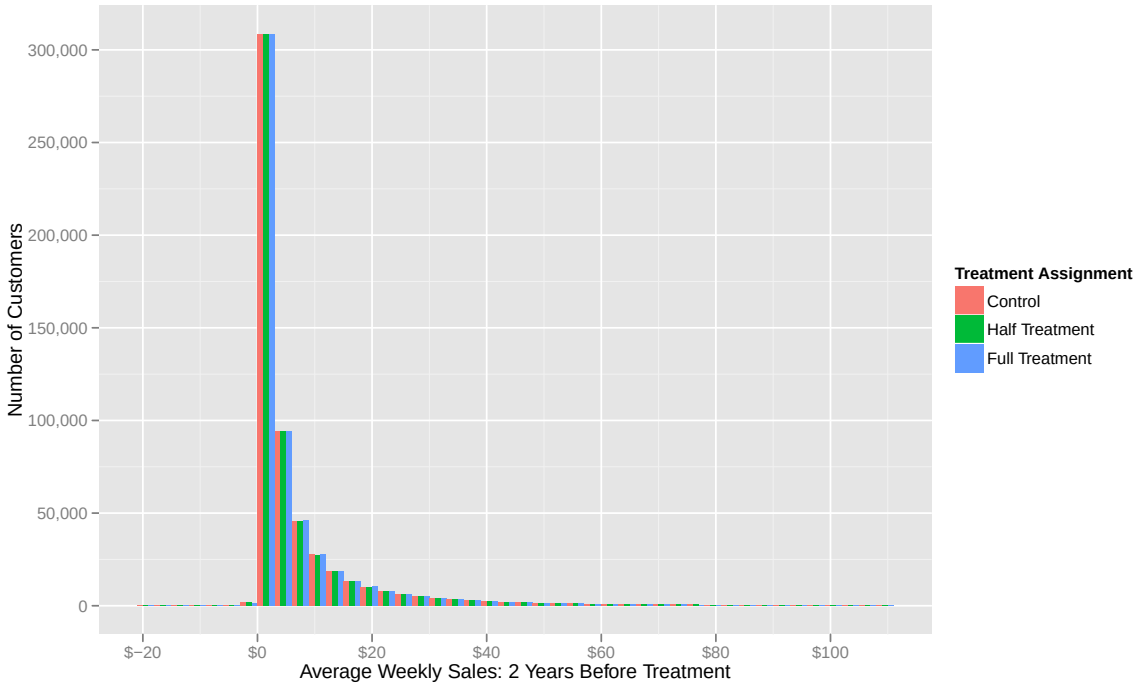
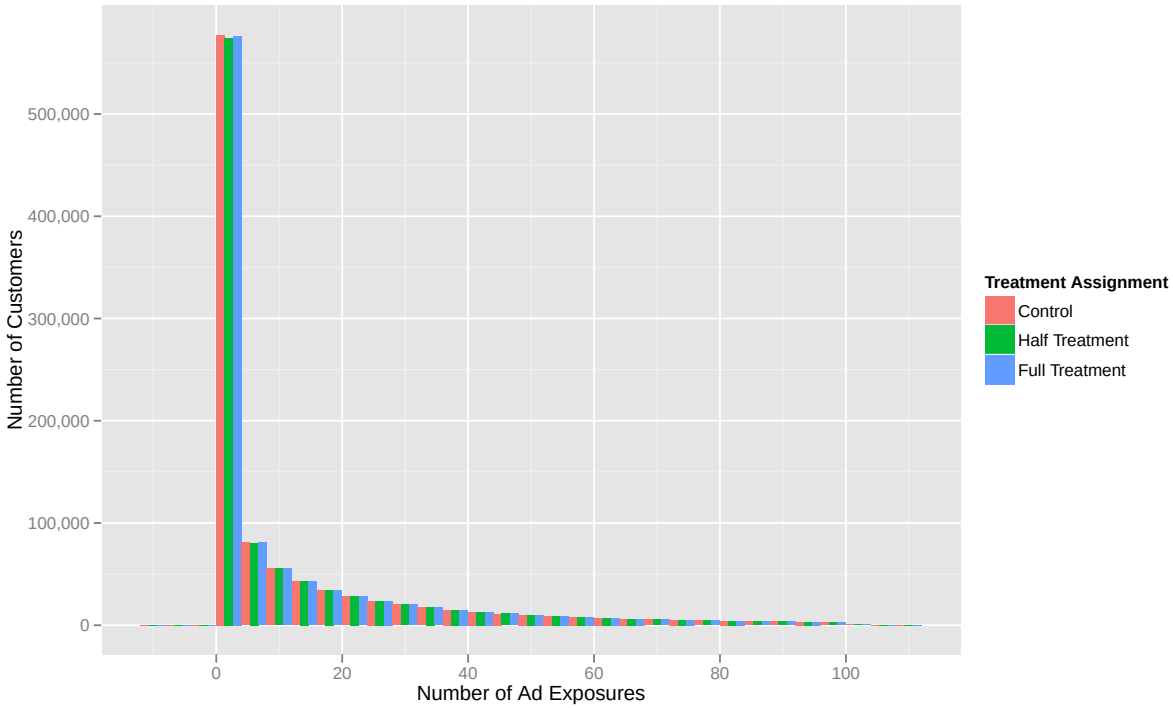


Figure 4: Histogram of Total Ad Exposures (Both Retailer and Control Ads)



A Online Appendix

A.1 Data Remarks

Our experiment resolves many traditional problems in measuring ad effectiveness. First, advertisers typically cannot identify the consumers who see their ads. We address this by restricting our experiment to logged-in, identifiable Yahoo! users. Second, advertisers rarely possess consumer-level data that links ad exposure to purchases. Our data are rare in that they combine sales data from the retailer—both online and in-store—with ad delivery and demographic data from Yahoo! at the consumer level.

We measure the effect of advertising on the retailer's relevant economic outcome—actual purchases—by relying on the retailer's customer-level data. The retailer believes that its data correctly attributes more than 90% of all purchases to individual customers by using all the information that they provide at check-out (credit-card numbers, phone numbers, etc.). We collect purchase data before, during, and after the campaigns.

We improve on the statistical precision of (Lewis and Reiley, 2014) by collecting both more granular sales data and sales data over a longer period of time. First, we obtain daily rather than weekly transactions during the ad campaigns. Daily transaction data allow us to discard purchases that take place prior to a customer's first ad exposure. Since pre-exposure transactions could not be influenced by the advertising, including such transactions in our treatment effect estimates only adds noise. This strategy avoids sample-selection problems, because the control ads identify the corresponding pre-treatment sales in the control group.[6] Second, we obtain a combination of weekly and aggregate consumer purchase data by channel for the two years prior to the experiment.[7] We use the purchase history as covariates in our TOT regressions to reduce the variance of

[6] If the ads affect behavior, this could create a selection effect that distorts the composition of the exposed sample or the number of ads delivered. Suppose that consumers are more likely to click on the retailer ad than the control ad. The ad-server may then have fewer opportunities to deliver ads because the people who click on the retailer ad are shopping rather than browsing Yahoo!. The summary statistics in Table 1 suggest, however, that ad exposure and browsing is sufficiently similar across groups that we can dismiss these concerns here.

[7] The data include weekly sales for the eight weeks before treatment. To save space, the retailer aggregated weeks 9–44 before treatment into a single variable. We have weekly data for weeks 45–60 before treatment, to capture any similar variation across years during the weeks of the experiment. The data again aggregate weeks 61–104 before

our experimental estimates.

We also use demographic data. Yahoo! requests user gender, age, and state at sign-up. A third-party data partner provided household income in five coarse strata.

Due to an unanticipated problem in randomly assigning treatment to multiple matched consumers, we exclude almost 170,000 users from our analysis. In particular, the third-party data collection firm joined 3,443,624 unique retailer identifiers with 3,263,875 unique Yahoo! identifiers; as a result, tens of thousands of Yahoo! identifiers were matched with multiple retail identifiers. The third party performed the experimental randomization on the retailer identifiers, but provided Yahoo! with only separate lists of Yahoo! identifiers for each treatment group to book the campaigns. Some multiple matched Yahoo! users were therefore accidentally booked into multiple treatment groups, which contaminated the experiment. To avoid this contamination, we discard all the Yahoo! identifiers who are matched with multiple retailer identifiers.[8] Fortunately, the treatment-group assignment is random, so the omitted consumers do not bias the experimental estimates. The remaining 3,096,577 uniquely matched Yahoo! users represent our experimental subjects. We acknowledge that our results only reflect 93% of exposed users.

Finally, we do not attempt to drop users with unusual browsing intensities. The maximum number of ad views in the experiment is 23,281, which we suspect is caused by automated software (i.e., a 'bot') running on that user's computer since the figure implies about 10,000 daily webpage visits. Though ads do not influence bots, we keep these observations in our analysis both because the appropriate cutoff is not obvious and because the upper tail of the distribution is small.

---

treatment into a single variable. Our data distinguishes between online and in-store sales throughout.

[8] We also perform the analysis on all uncontaminated customers assigned to a single group (results available from the authors upon request). We weight these customers to ensure the results represent the intended campaign audience. The re-weighting scheme increases the weight on multiple match consumers assigned to a single treatment. For example, a customer with three retailer identifiers who is assigned exclusively to the Full group receives a weight of nine in the regression, because uncontaminated customers represent three out of 27 possible combinations of triple treatment assignments. The results are qualitatively similar to those presented here, but statistically less precise. The weighted estimator has higher variance because the overweighted customers have higher variance in their purchases. For expositional clarity and statistical precision, we opt to discard multiple matched consumers here. Note that our point estimates of ad effectiveness are generally higher in the weighted analysis, so our preferred set of estimates are more conservative.

A.2 Channel, Campaign, Post-Campaign, & Shopping Trips Results

In this subsection, we collect results that decompose the ad effect by campaign, sales channel, shopping trips versus basket size and more. We use our preferred estimator from Section 5 throughout. This means that the regression model includes our full set of covariates and the outcome variable only includes purchases that take place after a consumer's first ad exposure.

A.2.1 Individual Campaign and Post-Campaign Impact

Table 3 considers the effect of advertising for both retailer ad campaigns individually and includes sales after the campaigns concluded.

The first two columns of Table 3 separately examine the two campaigns in the experiment. The two week-long campaigns are co-branded advertising that feature different clothing line brands. The point estimates for both treatment groups indicate that Campaign 2 is about three times more effective than Campaign 1, though the estimates from the two campaigns are not statistically distinguishable. Only the Full group during Campaign 2 demonstrates a statistically significant ad effect (p-value=0.012). Some of Campaign 2's success may be due to the lingering impact of Campaign 1, but we cannot test this hypothesis because we did not randomize treatment independently between campaigns.

The third column of Table 3 considers the lingering impact of advertising after the campaign concluded. To evaluate this, we use sales data from the two weeks after the campaign ended (the only post-campaign data available) and the total sales impact during and after the campaign. The point estimates for the Full and Half treatment groups indicate that the total campaign impact is respectively 10% and 64% larger when we include sales after the campaign. The total ad impact is marginally statistically significant for the Full group: \$0.525 for the Full group (p-value=0.089) and \$0.363 for the Half group (p-value=0.245). Note that the standard errors are higher than in our two-week estimates in Table 2, because the additional sales data increase the variance of the outcome variable. Since this increases the noise in our estimates more than the underlying signal, we treat these positive point estimates as suggestive of lingering effects.

These longer-term estimates allay somewhat the concern that the ad effect only reflects intertemporal substitution by consumers. If the ads simply cause consumers to make their intended future purchases in the present, then the short-run estimates will overstate the impact of advertising. In contrast, Simester et al. (2009) find evidence that short-run ad effects are due to intertemporal substitution among a catalog retailer's established customers. In an earlier experiment with the same retailer, Lewis and Reiley (2014) found a significant impact in the week after a two-week campaign and found suggestive evidence of an impact many weeks after this campaign.

A.2.2 Sales Channel: Online Versus In-Store

Table 4 decomposes the treatment effect into online versus in-store sales. The point estimate of the impact on in-store sales is \$0.323 for the Full treatment group, which represents 68% of the total impact of \$0.477 on sales repeated in column (1). The Half treatment group is similar as in-store sales represent 84% of the total treatment effect. These figures resemble the finding in the Lewis and Reiley (2014) experiment with the same retailer that found that in-store sales represented 85% of the total treatment effect.

We expect that online advertising complements the online sales channel: the consumer receives the ads when their opportunity cost of online shopping is lowest. Indeed, we find that—among control group members during the experiment—online sales are 11.5% higher among exposed users. Our Full group estimates suggest online sales increase by 6.8% over the Control group but in-store sales increase by only 3.0%. The proportional lift in the Half group is about the same: 1.6% for online sales and 1.7% for in-store sales.

A.2.3 Probability of Purchase Versus Basket Size

Marketers often decompose the effect of ads on sales into increasing the probability of purchase and buying more per shopping trip. We examine the experimental differences in probability of purchase, the number of shopping trips, and the 'basket size' or purchase amount conditional on a transaction. We present the basket size results as descriptive since we cannot separately identify

the basket size of marginal consumers (those for whom ad exposure caused them to make one or more purchases instead of zero) from those who would have made at least one purchase anyway. To examine the impact on shopping trips, we construct a variable equal to the number of days in which a consumer made a transaction during the campaign period.[9] We define this separately for online and in-store transactions and also sum these to get our measure of shopping trips as total transaction channel-days. For those customers who made at least one transaction in the two-campaign weeks, the mean number of channel-day transactions is 1.46.

Table 5 illustrates our results. The first column restates our original results for total sales. The second column presents results of a linear-probability regression for a transaction indicator dummy variable. The probability of a transaction increases with advertising by 0.43% (s.e. 0.46%) for the Full treatment group and by 0.47% (s.e. 0.46%) for the Half treatment group, relative to a baseline purchase amount of 7.7% for all treated consumers in the sample, though the increases are not statistically significant.[10]

Table 5's column (3) examines the impact on basket size. It restricts the sample to those 7.7% of consumers who made a transaction. The estimates suggest that the advertising increases the mean basket size by \$3.82 for the Full treatment group and \$1.27 for the Half treatment group, though neither of these coefficients are statistically significant. Relative to a baseline mean basket size of \$171, these represent percentage increases of 2.24% and 0.74% respectively.

Table 5's column (4) shows the impact of ads on shopping trips. The Full treatment produces 0.0020 additional trips (p=0.013) and the Half treatment produces 0.0011 additional trips (p=0.14) per person. These point estimates represent 1,092 incremental transactions in the Full group and 640 in the Half group. The additional columns of the table show that the effects are larger for in-store ($p < 0.1$) than for online sales ($p < 0.01$). Because the mean number of channel-day transactions per person is 0.112, the Full treatment effect represents a 1.8% increase in total transactions. This represents half of the 3.6% total treatment effect on sales. In contrast, Lewis and

[9] We define a transaction to be a net positive sale or negative return.

[10] This may surprise some readers who expect the statistical power problem in advertising to improve if we move from noisy sales data to a transaction indicator variable. Here, we see that the signal (0.0043 increase in transaction probability) is still two orders of magnitude smaller than the noise in transaction probability (s.d. 0.26).

Reiley (2014) found that increased probability of purchase represents only around one-quarter of the total effect on purchases. However, their data only allows them to examine the impact on the probability of any transaction during the campaign and misses the potential role of multiple shopping trips in the ad effect.

A.2.4 Difference-in-Differences and Observational Estimates

In Table 6, we compare the experimental treatment estimates with simple cross-sectional observational estimates and with observational estimates using the difference-in-differences approach from Lewis and Reiley (2014). In Table 6, we include average treatment-effect-on-the-treated estimates as a baseline. Since we are looking at observational approaches, we give estimates without including controls and without trimming pre-exposure outcomes—equivalent to Table 2, column (2). To better gauge the performance of the alternative estimators, Table 6 shows evidence for the effect of ads on total sales, in-store sales, online sales, and probability of purchase.

We first examine the biased cross-sectional estimates that we would get from observational data in the absence of an experiment. Table 6 provides cross-sectional observational estimates that difference treated users from the Full or Half treatment group with untreated users from the same group. For the total sales estimates, the cross-sectional observational point estimates are within a standard error of the experimental estimates. Closer inspection reveals this coincidence is mere luck and not generalizable, because five of the eight cross-sectional observational estimates in Table 6 exhibit significant biases.[11] The online sales observational estimates are much higher for both the Full ($0.512 vs. $0.197) and Half groups ($0.290 vs. $0.061). The greater estimate could be an due to activity bias (Lewis et al., 2011), where online behaviours—e.g., browsing on Yahoo! and purchasing online at the retailer—are correlated among users. The in-store sales cross-sectional estimate is lower for the Full group ($-0.160 vs $0.328), but about the same for the Half group ($0.114 vs. $0.127). The cross-sectional estimates for the probability of transaction lift are an order of magnitude larger than the experimental estimates. Collectively, this adds to evidence

[11] We test this formally by seeking to reject the hypothesis that the average outcomes among the untreated users in a treatment group the same as treated users in the Control group. Results are available from the authors by request.

from experimental studies that naive observational estimates exhibit biases (Gordon et al., 2015; Hoban and Bucklin, 2015; Lewis et al., 2011; Lewis and Reiley, 2014).

For our difference-in-differences estimates we use the same model as Lewis and Reiley (2014). That is, our outcome variable first-differences the outcome during the two week campaign and the outcome two weeks before the campaign. The regression then includes a term for treated users in the Full group and treated users in the Half group that takes a second difference between these groups and the untreated group of users drawn from the Full and Half groups and the entire Control group. Recall that Lewis and Reiley (2014) did not have control ads to distinguish the treated in the Control group, so used the difference-in-differences parallel-trends assumption to improve the precision of their estimates over Intent-to-Treat estimates. Table 6 shows that our difference-in-differences estimates line-up with our experimental differences estimates and the two estimates are within about a standard error of each other throughout. (However, we note that this DID estimator, used as the preferred estimator by Lewis and Reiley (2014), benefits from the exogenous variation created by the experiment. A fully observational study would have no exogenous variation available, and would instead rely fully on the endogenous variation between treated and untreated users.) With the exception of the online sales results, we find that the difference-in-differences lift point estimates are higher than their experimental counterparts.

In non-experimental settings, the difference-in-differences assumption of parallel trends between the treated and untreated cannot be verified. However, this assumption can be tested with an experimental control group. In our case, control ads allow us to directly compare the trends between the untreated in all three treatment groups with the treated users in the Control group. These tests reveal no significant differences in trends for any of the four outcome variables.[12] Without control ads, Lewis and Reiley (2014) still have a weaker test of the parallel trends assumption in which they compare the trends in the untreated Treatment group users with the entire Control group; they also can not reject that the trends are identical. Table 6's results may increase confidence in the difference-in-differences results in Lewis and Reiley (2014), but the results do not imply

[12]For total sales, this difference is \$0.256 (s.e. 0.290). For the remaining outcomes, the differences are respectively \$0.236 (0.254), \$0.020 (0.093) and -\$0.00013 (0.00053).

Table 3: Effects of the Advertising During and After the Campaign

	(1)	(2)	(3)
Timeframe	Campaign 1	Campaign 2	During & After Campaigns Total (4 weeks)
Subset of Users[a]	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x
Full Treatment ($)	0.116	0.382**	0.525*
	(0.144)	(0.153)	(0.309)
Half Treatment ($)	0.059	0.156	0.363
	(0.141)	(0.155)	(0.312)
Covariates: Full Set[c]	x	x	x
Observations	1,496,392	1,509,737	1,714,704
R^2	0.058	0.056	0.170

Average effect of Treatment on the Treated estimates. Dependent variable is sales during (or after) the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details).

that the difference-in-differences assumption is valid in their setting. In both cases, these tests are somewhat weak given the imprecise means of the outcome variable.

Table 4: Effects of the Advertising, Online versus Offline

	(1)	(2)	(3)
Dependent Variable	All Sales	In-Store Sales	Online Sales
Subset of Users[a]	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x
Full Treatment ($)	0.477**	0.323*	0.154**
	(0.204)	(0.172)	(0.0779)
Half Treatment ($)	0.221	0.185	0.036
	(0.209)	(0.176)	(0.081)
Covariates: Full Set[c]	x	x	x
Observations	1,714,704	1,714,704	1,714,704
R^2	0.091	0.078	0.135

Average effect of Treatment on the Treated estimates. Dependent variables are sales during the two weeks of the experiment. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. [a] Treated users are those who are exposed to either the retailer or the control ad. [b] Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c] Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details). We also include indicator variables for the given condition.

Table 5: Effects of the Advertising: Probability of Purchase, Basket Size, versus Shopping Trips

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Sales	Probability of Transaction	Sales Conditional On Transaction	Shopping Trips Online + In-Store	Shopping Trips In-Store	Shopping Trips Online
Subset of Users[a]	Treated	Treated	Transacted	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x	x	x	x
Full Treatment ($)	0.477**	0.000426	3.822	0.00196**	0.00129*	0.000662***
	(0.204)	(0.000461)	(2.391)	(0.000795)	(0.000698)	(0.000212)
Half Treatment ($)	0.221	0.000474	1.365	0.001162	0.000956	0.000205
	(0.209)	(0.000462)	(2.438)	(0.000793)	(0.000698)	(0.000211)
Covariates: Full Set[c]	x	x	x	x	x	x
Observations	1,714,704	1,714,704	132,568	1,714,704	1,714,704	1,714,704
R^2	0.091	0.147	0.107	0.174	0.171	0.090

Average effect of Treatment on the Treated estimates. Dependent variables are within the two weeks of the experiment. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
[a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details). We also include indicator variables for the given condition.

Table 6: Comparison with Difference-in-Differences & Observational Estimates

	(1)	(2)	(3)	(4)
Dependent Variable	All Sales	In-Store Sales	Online Sales	Probability of Transaction
Experimental estimates[a]				
Full Treatment	0.525**	0.328*	0.197**	0.000560
	(0.237)	(0.198)	(0.0914)	(0.000531)
Half Treatment	0.189	0.127	0.0613	0.000326
	(0.235)	(0.200)	(0.0880)	(0.000531)
Observational Estimates: Treated vs. Untreated[b]				
Full Treatment*	0.352	-0.160	0.512***	0.00485***
Treated	(0.254)	(0.217)	(0.0833)	(0.000555)
Half Treatment*	0.405*	0.114	0.290***	0.00467***
Treated	(0.243)	(0.211)	(0.0784)	(0.000554)
Difference-in-differences estimates[c]				
Full Treatment*	0.609**	0.442*	0.166**	0.000571
Treated	(0.274)	(0.241)	(0.0819)	(0.000507)
Half Treatment*	0.471*	0.418*	0.0531	0.000759
Treated	(0.280)	(0.242)	(0.0849)	(0.000507)

All estimates include no additional covariates, the dependent variables are outcomes during the two weeks of the experiment, and the dependent variables do not filter out pre-exposure activity. As in Table~2, we say a user in the Half group is treated if they are exposed to the retailer or control ad for simplicity. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]These are the baseline experimental estimates that replicate the analysis in Table 2, column (2). [b]The observational estimates run separate regressions for the Full and Half groups. For example, the Full group regressions reports the difference between the mean outcomes among treated users and untreated users. [c]The difference-in-difference estimates employ the model in Lewis and Reiley (2014). That is, the outcome variable first differences sales two weeks before and two weeks during the campaign and the second difference is between the treated users in the Full or Half groups and the untreated group users in all groups (including all users in the Control group).