Add More Ads? Experimentally Measuring Incremental Purchases Due To Increased Frequency of Online Display Advertising

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

April 27, 2013

First Draft: July 2011; This Draft: April 27, 2013

Abstract

Yahoo! Research partnered with a nationwide retailer to study the effectiveness of display advertising on online and in-store sales for more than three million shared customers. We measure the impact of higher ad impression frequency using a simple experimental design on Yahoo!: users in the ‘Full’ treatment group see the retailer’s ads, users in the ‘Control’ group see unrelated control ads, and users in the ‘Half’ treatment group see an equal probability mixture of the retailer and control ads. We find statistically significant evidence that the retailer ads increase sales 3.6% in the Full group relative to the control group. Doubling the average number of impressions per person, from 17 to 34 in a two-week period, nearly doubled the treatment effect. Leveraging our experimental design, we find that the returns to ad frequency are approximately linear among those who were eligible to see up to 50 ads and the marginal return to an additional ad exposure is 4¢. We also find intriguing evidence that the ads most strongly affected customers who live closest to the retailer’s brick-and-mortar locations, purchased recently, are loyal customers, and are wealthy.

*Johnson: Northwestern University, <garjoh@u.northwestern.edu>. Lewis: <randall@econinformatics.com>. Reiley: <david@davidreiley.com>. 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 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.
1 Introduction

Consumers often view the same advertising repeatedly. Most advertisers reasonably assume that diminishing returns eventually set in if the same consumer sees the ad enough times, but it is very difficult to know how many times is “enough” to produce these diminishing returns. In this paper, we shed light on this issue by running a large-scale controlled experiment on Yahoo!, with advertising exposure linked to individual retail purchases. Our field experiment examines the impact of a retailer’s advertising on customer purchases, both online and offline. We examine two consecutive weeklong ad campaigns that target the retailer’s existing customers. Our best experimental estimates suggest the retailer ads increased sales by 3.6% and that the campaigns were profitable. We also examine the impact of ad frequency by varying the intensity of treatment in the experiment. A minority of advertisers vary their demand for ads by consumer frequency and many of those that do set very low frequency caps. We show that the returns to advertising appear constant even out to 50 ad exposures in two weeks.

A report commissioned by the Interactive Advertising Bureau (Lavrakas, 2010) recently emphasized that controlled experiments are the only unbiased way to measure ad effectiveness, but such experiments remain quite rare in the industry. This first seems surprising, given the number of dollars at stake: $6.8 billion spent on online display advertising in 2011. However, experiments are costly to plan, operate, and analyze, even in the simple case of measuring outcomes via consumer surveys. These costs increase further when one wishes to measure the effect on the key economic outcome of actual consumer purchases, rather than proxy measures like click-through rate or increased brand recall. Furthermore, Lewis and Rao (2012) argue that the effect of advertising on purchases is statistically difficult to detect even when the true effect is economically large. They observe that the mean impact of advertising is plausibly a small fraction of the variance in purchases. Ad effectiveness studies therefore have low statistical power, and may require millions of observations to detect statistically significant effects.

Despite these challenges, experiments remain valuable relative to observational studies, which are prone to biases that undercut their usefulness. Observational studies estimate ad effects by comparing exposed users to unexposed users. Advertiser 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 and Reiley (2010) analyze an experiment in which a retailer targets its existing customers; they show that the observational estimate was four times the magnitude of the experimental estimate and had the wrong sign. Lewis, Rao and Reiley (2011) document a source of bias on the consumer side that they call ‘activity bias,’ which occurs because consumer Internet browsing varies across time and is correlated across websites. Lewis, Rao and Reiley (2011) show that activity bias leads to spurious overestimates of the causal effects of online advertising.

Our experimental design resolves many traditional problems in measuring ad effectiveness. First, advertisers typically can not 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. We use a third party company to link Yahoo! users to the retailer sales data. Our data are novel because they record not only the consumer’s online sales, but their in-store sales as well. Third, our study employs 3.3 million users in an effort to overcome the statistical power issue raised by Lewis and Rao (2012). In this large study, we could detect a campaign that returned a 50% profit on ad spend at the 10% significance level a disappointing 70% of the time in the Full group and 29% in the Half group. We improve this and increase the precision of our estimates by focusing the estimator on the sales of treated consumers after their first ad exposure, which we can identify using control ads and daily sales data.

We designed our experiment to examine the impact of ad 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 include a ‘Half’ treatment group that sees a mixture of the retailer and control ads with equal probability. This allows us to separate the impact of ad frequency from ‘ad type’, by which we mean the user’s eligibility to see ads based on her browsing intensity. We can then examine the shape of the consumer ad frequency response curve. Economists have debated whether the frequency curve is exhibits increasing, constant, or decreasing returns to scale.

Our results suggest that the retailer ads increased sales and that the campaign was profitable. Our preferred experimental estimates suggest an average impact of an additional ad to be $0.477 on the Full treatment group and $0.221 on the Half treatment group, which represents a 3.6% and a 1.7% lift over the Control group’s sales. The ads cost about 0.5¢ and the average ad exposure was 34 and 17 in the Full and Half groups respectively. Assuming a 50% retail markup, the ad campaign yielded a 51% profit but profitability is not statistically significant (p-value: 0.25, one-sided). The
lift in sales seems primarily due to in-store rather than online sales. We find weak evidence of heterogeneous treatment effects by consumer proximity to the retailer, purchase recency, loyalty, and wealth. Consumers exhibit large and significant ad effects when they live within one mile of the retailer, transact within eight weeks of the experiment, spend over $1000 in the past two years, or earn more than $100,000.

We find approximately constant returns to ad frequency for users who are eligible to see 50 ads or less. These users represent 80% of the treated population. We measure the marginal value of an additional ad to be 3.68¢ for this subpopulation. The experiment allows us to test several otherwise untestable assumptions about the impact of frequency. These assumptions include linear returns to ad frequency and no impact from ad type for users eligible to see up to 50 ads. Under these assumptions, our structural estimates for the marginal impact of an additional ad is nearly unchanged at 3.71¢. However, these assumptions increase the precision of our structural estimate. We can then show that the overall campaign was profitable at the 5% significance level exclusively using the sales lift among those consumers who were eligible to see up to 50 ads.

Our ad experiment is among the largest and most statistically powerful ever attempted. Split-cable television studies examine ad effectiveness using a panel of 3,000 viewers combined with data on consumer-packaged-good purchases (see Erdem and Keane (1996), and Lodish et al. (1995)). However, these studies have low statistical power. Meta-analyses of over 600 such tests by Lodish et al. (1995) and Hu et al. (2007) show that the majority failed to demonstrate statistically significant ad effects at the 5% level. Sahni (2012) examines the impact of display ads on a Indian restaurant search website with eleven separate experiments involving between 3,000 and 53,000 users. These experiments are statistically powerful enough to measure ad effectiveness because they target consumers who are in the market to find a restaurant and use 'sales leads' (clicking for restaurant number) as the outcome variable rather than sales. We improve on Lewis and Reiley (2010)’s previous experiment with the same retailer by doubling the sample size from 1.6 million customers and improving the experimental design—notably by including control ads.

Previous authors have considered the relationship between ad frequency and effectiveness. Much ad frequency research relies on observational studies (see Pechmann and Stewart (1988) for an early survey). Our results stand in contrast to Kameya and Znija (2002)’s meta-analysis of observational online display ad frequency effects which found returns are diminishing after only five impressions.
Several recent papers use experiments to examine ad frequency. Lewis (2010) examines the impact of frequency of online display ads on clicks and conversions using natural experiments reaching millions of Yahoo! users. He finds heterogeneous frequency effects among campaigns: some display decreasing returns and others display constant returns up to 50 ad views. Anderson and Simester (2004) experimentally vary the frequency of a catalog retailer’s mailings with a sample of 20,000 customers. Though more mailings increased short-run purchases, they attribute this in part to inter-temporal substitution especially among the firm’s best customers. Sahni (2013) also varies treatment intensity in his Indian restaurant display ad experiment. His estimates suggest decreasing returns even after three exposures.

The rest of this paper is organized as follows. The next section describes our experimental design and data collection. The third section gives an overview of the data and its descriptive statistics. The fourth section presents our measurements of the causal effects of the advertising, particularly about the effects of frequency, as well as heterogeneous treatment effects by consumer characteristics. The final section concludes.

2 Experimental Design

The experiment measures ad effectiveness for a national apparel retailer advertising on Yahoo!. The experiment took place during two consecutive weeks of 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 implemented 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. 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 experiment employs control ads to identify the counterfactual ad exposures that a consumer could have seen if they were assigned to the full retailer campaign. The control ads also identify the counterfactual treated subpopulation so that we can measure more precisely the average treatment effect on the treated (ATET). Without control ads, Lewis and Reiley (2010) compute the ATET indirectly by differencing the average outcomes of consumers assigned to treatment from the average of those assigned to the control group (intention to treat) then dividing by the probability of receiving treatment. Though the two measures of ATET are equivalent, the indirect measure is less precise because it includes noise from the difference in outcomes among the untreated subpopulations, which is asymptotically zero.

The experimental subjects are existing customers of both Yahoo! and the retailer. Prior to the experiment, a third-party data-collection firm created a list of joint customers who matched on both name and address (either terrestrial or email). The third-party firm doubled the list of joint customers from the 1.6 million customers studied by Lewis and Reiley (2010) to the 3.3 million customers in the present experiment by leveraging the third-party firm’s own customer addresses record. After the experiment ended, the third-party firm combined the retailer sales data and the Yahoo! advertising data. The third-party firm removed identifying information including names and addresses in order to protect customer privacy.

Due to an unanticipated problem in randomly assigning treatment to multiple matched consumers, we exclude almost 170,000 customers from our analysis. We engaged the third-party data collection firm to create a list of experimental subjects from the Yahoo! and retailer customer lists. The third-party firm’s data contained 3,443,624 unique retailer identifiers but 3,263,875 unique Yahoo! identifiers, as 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! only with 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. 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 the subjects examined the experiment. We acknowledge that our results only reflect 93% of the original population.

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 on attractive models. The creative content of each ad impression is dynamic and involves slideshow-style transitions between still photographs and text. Campaign 1 includes three different types of 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. The Yahoo! Search ad demonstrates an animated search for the rock band ‘The Black-Eyed Peas’ and emphasizes both the search links and multimedia content that Yahoo! Search provides. The animation ends with an in-ad, functioning search box. The control ads are the same for both weeks.

The experiment’s ads appear on all Yahoo! properties, such as Mail, News, Autos, and Shine. The ads take four rectangular formats, which appears different Yahoo! pages. We ensure the distribution of ad formats is the same for all three treatment groups.

3 Data

The data section contains three subsections. The first describes our data and demonstrates that our experimental randomization is valid. The second presents a power calculation that shows the difficulty of measuring ad effectiveness in our setting. The third section details the source of our data and explains how we generate some of our key variables.

\[1\text{We also perform the analysis on all uncontaminated customers assigned to a single group (results available from the authors upon request). We weight these customers to ensure the results represent the intended campaign audience. The re-weighting scheme increases the weight on multiple match consumers assigned to a single treatment. For example, a customer with three retailer identifiers who is assigned exclusively to the Full group receives a weight of nine in the regression, because uncontaminated customers represent three out of 27 possible combinations of triple treatment assignments. The results are qualitatively similar to those presented here, but statistically less precise. The weighted estimator has higher variance because the overweighted customers have higher variance in their purchases. For expositional clarity and statistical precision, we opt to discard multiple matched consumers here. Note that our point estimates of ad effectiveness are generally higher in the weighted analysis, so our preferred set of estimates are more conservative.}\]
3.1 Descriptive Statistics

Table 1 provides summary statistics for the experiment. Over three million customers were evenly assigned to one of our three treatment groups: Full, Half, or Control. 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.

The summary statistics in Table 1 are consistent with a valid experimental randomization. In each treatment group, approximately 68.5% of customers are female, the average age is 43.6 years, and customers viewed an average of 122 webpages 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 1 shows that the consumer proximity’s to the nearest retailer is distributed near identically across treatment groups. Treated consumers live on average 21.3 miles from the nearest retailer location and the median proximity is 6.6 miles.

In Table 1, we see that the experiment delivers advertisements evenly across treatment groups. 48.3% of users are exposed to Campaign 1, 48.8% are exposed to Campaign 2, and 55.4% see ads in at least one campaign. Campaign 1 delivers 13% more ad impressions per person than Campaign 2, but reaches 1% fewer customers. For the remainder of this section, our descriptive statistics exclude the unexposed users. Figure 2 shows that the distribution of total ad views (both retailer and control) across the three treatments is identical. The distribution of ad views is highly skewed towards the left, so that the mean is 33 while the median is 15. As expected, the Half treatment group sees an even split of retailer and control ads. On average, the exposed subjects saw an experimental ad on 16.3% of the pages they visited on Yahoo! The maximum number of ad views was 23,281, which we suspect was caused by automated software (i.e., a ‘bot’) running on that user’s computer since the figure implies about daily 10,000 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.

---

2 Customer age is self-reported by users to Yahoo! and has proven to be reliable. However, outliers and miscoded values remain for a small fraction of users. To guard against the effects of outliers, we bottom-code age at 14 years and round up the reported age of 0.16% of the population. Similarly, we top-code age at 86 years and round down the age of 0.56% of the population.

3 An F-test for the null hypothesis of equality of the means between the three treatment groups gives p-values for these five variables of 0.794, 0.607, 0.132, 0.485, and 0.517 respectively.

4 Users who see at least one ad visit an average of 205.6 pages on Yahoo! during the two week experiment.

5 A few extreme outliers cause the page views and ad views to differ across treatment groups in the third decimal place.
Table 1 also describes clicks on the retailer’s ads. The click-through rate—the quotient of clicks and ad views—is 0.17% for Campaign 1 and 0.22% for Campaign 2 in the Full group. These click-through rates are high: many online display advertising campaigns have click-through rates under 0.1% Lewis (2010). The clicker rate—the fraction of exposed users who click on any ad—is 2.9% for Campaign 1 and 3.2% for Campaign 2 in the Full group. The Half treatment group’s clicker rates are lower because its subjects have fewer opportunities to click an ad.

Figures 3 and 4 show that the distribution of pre-treatment sales are essentially identical across all treatment groups. Figure 3 illustrates the distribution of average weekly sales over the two years prior to the experiment. Figure 4 shows the distribution of individual purchases conditional on transactions in the two weeks prior to the experiment. To better visualize the distribution, Figure 4 omits the 90.8% of observations with zero purchases in the two weeks. 0.96% of purchase amounts are negative, which represents returns to the store net of any purchases.

Figure 5 shows sales for the three treatment groups during the two-week experiment, excluding customers with zero purchases. Figure 5 shows how the effect of advertising is small relative to overall sales even though the estimated lift is economically substantial relative to the ad cost. We can distinguish a slight shift in the distribution to higher levels of purchases: negative net purchase amounts are less frequent in the treatment groups than the control group, while purchase amounts larger than $200 appear more frequent in the treatment groups. We see that we are at the measurement frontier for detecting the treatment effect of ads on sales.

### 3.2 Power Calculation

Since our sample is much larger than most experimental studies, we wish to temper the reader’s expectations regarding the strength of our experimental results. The experiment’s statistical power is limited even though our experiment includes three balanced treatment groups with about 570,000 treated users each. For a comprehensive discussion of the statistical power issue, we refer the reader to Lewis and Rao (2012)’s meta-analysis of over twenty experiments measuring online display ad effectiveness.

To demonstrate the limits of our experiment, we present a statistical power calculation for testing the null hypothesis that advertising has no impact on sales. In the calculation, we consider place, contributing to the relatively low $p$-value mentioned for page views in footnote 3.
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% gross margin for the retailer.
That is, the null and alternative hypotheses are

\[ H_0 : \Delta sales = 0 \]
\[ H_a : \Delta sales = 0.51, \]

To determine power, we first calculate the expected \( t \)-statistic under this alternative. 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. The expected \( t \)-statistic is given by

\[
E[T] = \frac{\hat{\delta} - 0}{SE(\hat{\delta})} = \frac{0.51 - 0}{\frac{125\sqrt{2}}{\sqrt{570,000}}} = 2.18
\]

for the average treatment effect estimator \( \hat{\delta} \).\(^6\) Using the 10% one-sided critical value of \( t^* = 1.645 \),
the power of the test is given by

\[ Pr(T > t^*|E[T] = 2.18) = 70\% \]

An analogous test for the Half group only has power of 29%.\(^7\) We emphasize that the above calcu-
lations are about whether the ads impact sales: a test of profitability has much less power because
it requires gross profits to exceed positive costs, rather than zero. The above power calculations
elucidate both the limits of our experimental results and the importance of improving the precision
of our treatment effect estimates. In Section 4.1, we describe several econometric methods designed
to improve precision.

\(^6\)Let \( \bar{\mu} \) denote the sample average and \( \sigma^2 \) the variance of \( N \) consumer sales observations. Let the subscripts \( T \) and
\( C \) denote the Treatment and Control groups. The standard error of the treatment effect estimator \( \hat{\delta} \) is given by the
square root of its variance

\[
Var(\hat{\delta}) = Var(\bar{\mu}_T - \bar{\mu}_C) = \frac{\sigma^2_T}{N_T} + \frac{\sigma^2_C}{N_C} = \frac{2\sigma^2}{N}
\]

where we assume that \( N_T = N_C \) and \( \sigma^2_T = \sigma^2_C \), which is a good approximation here.

\(^7\)The alternative hypothesis for the Half group effect is \( \Delta sales = 0.255 \) with \( E[T] = \frac{0.255 - 0}{\frac{125\sqrt{2}}{\sqrt{570,000}}} = 1.09 \). The
associated power is \( Pr(T > t^*|E[T] = 1.09) = 29\% \).
3.3 Data Source

Our data combines sales data from the retailer with ad delivery and consumer demographic data from Yahoo!. Our data are rare in that they not only match ad exposure and sales at the consumer level, but also record both online and in-store sales.

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 our predecessor (Lewis and Reiley, 2010) by collecting a combination of 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 counterfactual pre-treatment sales in the control group.\(^8\) Second, we obtain consumer purchase data for the two years prior to the experiment.\(^9\) We use the purchase history as covariates in our ATET regressions to reduce the variance of our experimental estimates.

We also examine how the effect of advertising varies with a customer’s proximity to the nearest retailer store. To compute proximity, we use nine-digit zip code data from Yahoo! and a third-party data broker, which are available for 75.2% of the experimental subjects and 77.8% of the exposed subjects. A nine-digit zip code essentially indicates a city block, which provides the fine-grained resolution required to investigate the effect of advertising within a mile of a store. We employ a database from Yahoo! Maps to link nine-digit zip codes to latitude and longitude coordinates.\(^10\)

---

\(^8\) 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.

\(^9\) 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 treatment into a single variable. Our data distinguishes between online and in-store sales throughout.

\(^10\) Two issues arose with the zip code data: 1) the Yahoo! Maps database did not have zip code coordinates for
For each customer, we compute the ‘crow-flies’ distance to the nearest store using the haversine formula.\textsuperscript{11}

We also use demographic data from Yahoo!. Yahoo! requests user gender, age, and state at sign-up. We use income data from Yahoo!’s behavioural targeting database. This data comes from coordinating with third party data brokers. The data is coarse and consists of five income strata. We use the most recent available measurements of the three available measurement dates.

4 Results

We find significant evidence that the experiment’s advertising causes an increase in consumer sales. Our preferred estimates suggest that consumers who were exposed to the retailer’s ad saw their average sales increase by \$0.477 with standard error (s.e.) \$0.204 in the Full treatment group and \$0.221 with s.e. \$0.209 in the Half treatment group. These represent a 3.6% and 1.7% increase in sales above the Control group. The effect in the Full group is highly significant (\(p\)-value, two-sided: 0.020), the Half group is not significant (\(p\)-value: 0.289), and the joint test is marginally significant (\(p\)-value: 0.065). Though the effect in the Half group is 46.3% that of the Full group, an F-test for equality of the two effects is not rejected at 10% (\(p\)-value: 0.211).

We also leverage the experimental variation in the consumer’s retailer ad views (frequency) and eligibility for ads (ad type) to measure the average impact of an additional retailer ad. We restrict our analysis here to the 80% of consumers who see up to 50 total ads (retailer and control) for whom our estimates are more precise. Our pure experimental measure suggests that an additional ad has a 3.68\(\cent\) impact (s.e. 1.72\(\cent\)). We also use the experiment to evaluate several structural assumptions that relate ad effectiveness to frequency and ad type. Under these assumptions, our structural

\textsuperscript{11}The haversine formula calculates the distance between two pairs of latitude and longitude coordinates while correcting for the spherical curvature of the planet.
estimates suggest the value of an additional ad is almost unchanged at 3.71¢ but more precisely estimated with standard error of 1.02¢.

Our experimental evidence suggests that the ads were likely profitable for the retailer. Given about 570,000 exposed users in each of the three treatment groups, our preferred experimental estimates and 95% confidence interval imply a total sales lift of $273,000 ± 229,000 in the Full group and $126,000 ± 234,000 in the Half group. The campaign costs about $88,000 for the Full campaign and $44,000 for the Half campaign. To calculate profitability, we assuming a conservative gross margins of 50% for the retailer. Our point estimates indicate a rate of return of around 50% on the advertising dollars spent but the 95% confidence intervals extend from losing the entire investment to doubling it.

We consider heterogeneous treatments by consumer characteristics. We find evidence of heterogeneous response by consumer proximity to the retailer, purchase recency, retailer loyalty, and income. We find large and statistically significant ad effects within the subpopulations who live within one mile of a retailer, transacted within eight weeks, spend more than $1,000 at the retailer in the previous two years, earn more than $100,000 or are female.

We further decompose the result by campaign, sales channel, and shopping trips. Our estimates suggest the second weeklong campaign is more effective and demonstrates a statistically significant ad effect on its own (p-value: 0.012). We find that the majority of the total treatment effect is attributable to the in-store rather than online sales channel: 68% in the Full group and 84% in the Half group. We present descriptive evidence that decomposes the treatment effect into increased probability of purchase versus basket size. Though these results are inconclusive, we are able to detect a statistically significant 1.8% increase in shopping trips in the Full group.

The results section is divided into four subsections. The first shows 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. The second subsection analyzes the role of ad frequency on ad effectiveness. We first measure this using experimental variation then present structural estimates based on assumptions we show are supported by the data. The third section examines heterogeneous treatment effects by consumer characteristics. The final subsection collects additional results that break apart the effect of advertising by online versus in-store sales, by individual ad campaign, by probability of purchase versus basket size.
4.1 Overall Campaign Impact

Table 2 presents regression estimates of the Average Treatment Effect on the Treated (ATET) for
the impact of advertising on consumer purchases during the two-week experiment. In particular,
Table 2 highlights the various methods for increasing the precision of the estimates. These methods
include pruning components of the outcome data that can not be influenced by advertising and
introducing covariates into the regression. In all, we improve the precision of our estimates—or
shrink the standard errors—by 34% on average.

Table 2 begins with the indirect ATET estimate. The indirect ATET estimator takes the
treatment-control differences for the entire sample of 3.1 million consumers (intent to treat estimate)
and divides by the treatment probability (the 55.4% exposed subsample). The indirect ATET
estimator relies on the fact that outcomes among untreated subjects have an experimental difference
of zero asymptotically. In small samples, the difference among untreated subjects adds noise to the
estimator however. The indirect ATET estimator yields a $0.67 average sales lift (s.e. $0.32) in
the Full treatment group with and an average lift of $0.03 (s.e. $0.31) in the Half group. Whereas
Lewis and Reiley (2010) estimate the ATET indirectly out of necessity, this experiment employs
control ads to identify the counterfactual treated sample in the Control group.

Table 2’s column 2 presents the direct ATET estimate on the treated subsample. The direct
ATET estimator increases precision by pruning the untreated subsample which contributes only
noise to the estimator. The regression in column two shows that the Control group has $15.53
purchases on average while the average purchases in the Full treatment group are $0.52 larger
and those in the Half group are $0.19 larger. The Full treatment effect is statistically significant
($p=0.027$, two-sided), while the Half treatment is not ($p=0.423$). An F-test of joint significance
is marginally significant ($p=0.082$). The direct ATET requires control ads to identify the treated
subpopulation in the Control group, which improves the precision of the estimates by 25% on
average.

Table 2’s column 3 uses both the control ads and daily level sales data to further prune the data
and boost precision by another 8%. Specifically, we omit purchases that occur prior to a consumer’s
first experimental ad exposure. This method is free from bias because ads can not influence sales

---

12 This is numerically equivalent to computing a local average treatment effect by using the randomized intent
to treat as an instrument for actual treatment.
until the user receives the ad and the control ads identify the counterfactual pre-treatment sales in the Control group. Excluding irrelevant sales reduces the baseline average purchase amount from $15.53 to $13.17 per person. The point estimates here are $0.56 (s.e.: $0.22) and $0.31 (s.e.: $0.22) in the Full and Half groups.

In columns 4 to 7 of Table 2, we increase the precision of our estimates by adding covariates to the ATET regression. The covariates in the regression improve precision by reducing the residual, or unexplained, variance in the dependent variable. First, we add demographic covariates that 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!. These covariates explain little of the variance in sales since the $R^2$ is 0.001, so the precision of the estimates is unchanged. The fifth column adds retailer-defined customer segments: five categories for recency of last purchase, four categories for frequency of past purchases, and four categories for total lifetime spending at the retailer. The customer segments improve the $R^2$ to 0.042, and increase precision by 2%. The sixth column adds two years of individual-level past sales data and pre-treatment sales during the campaign, separately for online and offline sales. Past purchases best explain the sales during the experiment as they increase the $R^2$ to 0.090, and the standard error for the Full treatment estimate falls a further 3% on average. Table 2’s column 7 presents our preferred estimates, which include browsing intensity covariates. Specifically, we include fixed effects for the total experimental ads delivered (ad type) and indicators for the day of the consumer’s first ad exposure. To the extent that shopping behavior is correlated with current online browsing activity, the ad type fixed effects will improve efficiency. These covariates marginally improve the $R^2$ to 0.910 and decrease the point estimates without noticeably increasing precision.

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. However, covariates may be inexpensive to include and serve to validate the experimental randomization. Control ads are expensive, but they facilitate data pruning and thereby improve precision five time more than our covariates.

---

13Sales data provided at the level of individual weeks or aggregates of multiple weeks. See footnote 9.
14The ad type covariates include indicator variables for ad types (total ad views) from 1 to 29 and a single indicator variable for ad types of 30 or more.
15Note that, if we include the non-ad covariates in the indirect ATET estimator (column 1) to show the value of covariates alone, we still get a 5% improvement in precision.
We know an econometric strategy that would further improve precision, but we omit this for
simplicity and brevity. In essence, we can use the Half group members who receive a single treatment
(either retailer or control ad) to improve our estimates of the Full and Control treatment group
means. Conversely, we can also use the Full and Control groups to better estimate the Half group
mean outcomes. We would need to weight the re-assigned users appropriately to prevent bias.
The method also indicates that our Half group’s average treatment estimates are diluted by the
those who see only control ads and the Half group’s first experimental impression—used in the
trimmed outcome variable—could be either a retailer or control ad. Nonetheless, we do not pursue
this strategy in the interest of expositional simplicity and because this method is specific to an
experiment with partial treatment.

Our preferred estimator—in column seven 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.
The Full treatment effect is statistically significant at the 5% level (p-value: 0.020) and represents a
nearly 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 to have much confidence in the result. In the next
subsection, we use consumer-level variance in treatment intensity to better model the effect of ad
frequency.

Given 570,000 exposed users in each of the three treatment groups, our point estimates indicate
that the Full campaign increased retail purchases by a dollar value of $273,000 ± 229,000, while
the Half campaign increased purchases by $126,000 ± 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 indicate incremental revenues of around three times cost.\textsuperscript{16} We assume a conservative gross
margin of 50% for the retailer’s sales and find that our point estimates indicate a rate of return of
51% on the advertising dollars spent, but with 95% confidence intervals of [-101%, 204%].

\textsuperscript{16}Note our cost figures exclude ads delivered to the 7% of users whom we removed from our analysis due to the
problem of multiple matches. We have scaled the cost of the campaign proportionally based on the 92.8% included
share of total impressions.
4.2 Effects of Frequency

In this section, we answer our title’s question: should the advertiser “add more ads?” Specifically, we answer the question: what is the marginal value of an additional ad impression? Above, we learned that doubling the number of ad impressions to the treatment group increased the estimated treatment effect from $0.477 to $0.221 per person. However, the confidence intervals on these point estimates are sufficiently wide that we can not reject the hypotheses that doubling the number of ad exposures either 1) doubles the sales lift; or 2) has no additional effect. We now exploit consumer-level variation in ad frequency to increase the precision of our estimates.

We want to distinguish between the advertiser’s decision to purchase more ads and the user’s decision to visit more webpages on Yahoo!. Both choices increase the chance that a user sees more ads, but the advertiser can only impact their own ad intensity. Towards this, we introduce the twin concepts of ad frequency and ad type. We define a user’s ad frequency—denoted by $f$—to mean her number of retailer ad exposures. We define a user’s ad type—denoted by $\theta$—to mean her potential ad exposure given by the sum of her control and retailer ad exposures. The user’s browsing intensity determines her ad type: ad type is therefore endogenous.

The data contain three sources of variation in ad frequency. First, the three treatment groups provide exogenous variation in frequency that we exploit in Section 4.1. Second, the Half treatment group provides exogenous variation in frequency within ad types. Since each ad exposure has a 50% chance of being the retailer ad, the number of retailer ad exposures is binomially distributed, $f \sim \text{Binomial}(0.5, \theta)$. Lewis (2010) exploits binomial variation alone to assess the effects of frequency on the probability of clicking at least one ad. The variation in $f$ within the Half group is less than 5% as large as the variation between treatment groups; intuitively, the number of retailer ad exposures clusters fairly tightly around $\frac{1}{2} \theta$ for users with $\theta \geq 10$.\(^{17}\) Third, user-level variation in ad type $\theta$—due to browsing intensity—provides variation in $f$. Unlike the other two sources of the variation in $f$, the variation due to $\theta$ depends endogenously on user behaviour.

Our frequency estimates concentrate on exposed users who were eligible to see up to 50 experimental ads during the two-week campaign, $0 < \theta \leq 50$. This range includes 80% of treated subjects.

\(^{17}\)The variance of the binomial distribution is $Np(1 - p)$ while $N$ times the variance of a single Bernoulli trial (assignment to the Full treatment or control group) is $N^2p(1 - p)$. Hence, the ratio of variances is $1/N$. In terms of total variation, however, we have twice as many observations in the Full treatment and Control, hence the total variation in the Half group is half the combined variation of the Full treatment and Control groups.
Data on users who see more than 50 impressions is much more sparse, with some of these customers seeing hundreds or even thousands of impressions during the two-week campaign. Our estimates of the effects of frequency for these users are therefore imprecise. In a linear regression, the outliers have sufficient leverage to eclipse the measured effect on the majority of customers. Nonetheless, we show estimates for $\theta > 50$ below.

Our frequency analysis subsection contains two sections that present our experimental and structural estimates for the marginal value of an additional retailer ad. In the first section, we show, in the context of a very general model, that the average value of an ad impression is $3.68\$ (s.e. $1.72\$) among exposed users with an ad type of at most 50. In the second section, we make structural assumptions on the value of an additional impression, which we verify using the experiment. Under these assumptions, we find that the value of an additional impression is essentially unchanged at $3.71\$ (s.e. $1.02\$) for $0 < \theta \leq 50$, but the precision of our estimates increases by 41%.

### 4.2.1 Experimental Estimates

We present our most general experimental frequency effect estimates here. These rely purely on exogenous variation. Here, we estimate the model on exposed users, but we do not include covariates or restrict the outcome variable to post-exposure sales for the sake of brevity and clarity. Unfortunately, we lack statistical power to say much in the most general model even in more dense data for the ad type $0 < \theta \leq 50$.

Ideally, we would estimate mean purchases as a completely nonparametric function of ad frequency $f$ and ad type $\theta$. Letting $Y$ denote purchases and $\varepsilon$ denote the error term, we write the fully general model as

$$Y = g(f, \theta) + \varepsilon$$

However, we have insufficient statistical power to meaningfully estimate the fully general model because we have too few observations. Furthermore, our experimental variation in $f$ is non-uniform: the three treatment groups generate observations at $f = 0$ in the Control group, $f = \theta$ in the Full group, and centered around $f = \frac{1}{2} \theta$ but binomially distributed in the Half group.
As a point of departure, we instead propose a general model from Lewis (2010)

\[ Y = \beta(\theta) \cdot h(f) + \gamma(\theta) + \epsilon \]

The model proposes that \( \theta \) and \( f \) are multiplicatively separable for \( f > 0 \). Since \( f = 0 \) has no effect on sales, \( \gamma(\theta) \) captures the relationship between baseline sales and the endogenous \( \theta \). \( \beta(\theta) \) captures how the average effect of frequency varies with the potentially endogenous \( \theta \). \( h(f) \) captures how sales vary proportionally with frequency. We are primarily interested in the function \( h(\cdot) \) to determine the value to the advertiser of purchasing more ads.

In our experimental estimates, we assume that \( h(f) \) is linear, but allow \( \beta(\theta) \) and \( \gamma(\theta) \) to be fully non-parametric. We can therefore rewrite the above equation as

\[ Y = \beta_\theta \cdot f + \gamma_\theta + \epsilon \]

where \( \beta_\theta \) and \( \gamma_\theta \) represent a vector of 50 dummy variables for \( 0 < \theta \leq \theta \). By assuming that \( h(f) \) is linear, \( \beta_\theta \) can be interpreted as the marginal value of a retailer ad \( f \) for a given \( \theta \). In Section 4.2.2, we reexamine this assumption. The model can be estimated using a single fixed effects regression on 100 variables (50 fixed effects and 50 interactions of those fixed effects with \( f \)), or equivalently using 50 separate regressions of sales on \( f \) for a given \( \theta \).

Figure 7 presents all 50 estimates of the marginal impact of an additional ad \( \beta_\theta \) to users of type \( \theta \). The estimates are noisy, but generally positive. Though lower \( \theta \) have more observations, the confidence intervals are tighter for higher \( \theta \) because they contain greater variation in \( f \) that increases leverage under the linearity assumption. Table 3 provides an average marginal ad effect estimate of 3.68¢ (s.e. 1.72¢) across all \( \beta_\theta \) that is weighted by retailer ad exposures.\(^{18}\) For this ad campaign, the retailer paid an average cost of 0.46¢ per impression, so the marginal revenue of an additional impression is 8 times the marginal cost for those consumers with \( 0 \leq \theta \leq 50 \). In industry parlance, this is an 800% return on ad spend for the consumers we are studying.

We consider the impact of ad frequency for the 20% of users with ad type \( \theta > 50 \). We begin by computing the overall treatment effect using our preferred specification from Section 4.1, but

\[ 18 \hat{\beta} = \sum w_\theta \hat{\beta}_\theta \text{ and } Var(\hat{\beta}) = w_\theta^2 Var(\hat{\beta}_\theta) \text{ where } w_\theta = \frac{\sum w_i f_i}{\sum f_i}. \]
aggregating the Half and Full groups into a single treatment group. We obtain a noisy average treatment effect estimate of -$0.058 (s.e. $0.483) for the subset with $\theta > 50$. The subjects with $\theta > 50$ have an average ad frequency $f$ value of 88. We then divide the 95% confidence interval’s upper bound of the treatment effect $0.89$ by 88, which yields $0.01$ as an upper bound on the average treatment effect per ad exposure for users with $\theta > 50$. The estimated average effect of frequency is thus lower for $\theta > 50$ than our estimate of $0.037$ for $0 < \theta \leq 50$. The low marginal ad effect for high $\theta$ could either be due to diminishing marginal returns to frequency or due to lower ad effectiveness for high $\theta$ users. Since $f$ and $\theta$ are highly correlated, we cannot distinguish between the two explanations.

Regardless, the advertiser would want to cap the number of ads per person somewhere on the order of 50 ads over two weeks—provided the cap is not too expensive. The optimal frequency cap is much larger than the weekly caps of about 3 online display ads that we have often seen in practice.

4.2.2 Structural Estimates

We now propose and test several structural assumptions that improve the precision of our estimates. We will show that we are able to do so without introducing substantial bias. The reader may be surprised that an experimental paper includes a structural model. While we concede that experimentalists generally resist making strong assumptions, we argue that this resistance focuses on untestable assumptions. However, here we can leverage our experiment to test our assumptions. Below, we describe our assumptions which restrict the relationship between ad frequency $f$ and ad type $\theta$ and the relationship between baseline sales and ad type for $0 < \theta \leq 50$.

**Assumption 1.** a) $\beta(\theta)$ is constant; and b) $h(f)$ is linear, which imply

$$Y = \beta \cdot f + \gamma_\theta + \epsilon \quad (2)$$

Assumption 1a) assumes a constant marginal effect across ad type $\theta$ for the effect of frequency $f$ on sales. Assumption 1a) can also be viewed as the average treatment effect across heterogeneous effects. This seems reasonable because the retailer-ad weighted average of the estimated $\beta_\theta$ 3.68¢ lies within the 95% confidence intervals of nearly all the individual $\beta_\theta$ estimates in Figure 7. Moreover, the $\beta_\theta$ estimates in Figure X exhibit no clear trend that would challenge this assumption. We will
reexamine Assumption 1b)—the linearity of \( h(f) \) maintained from the experimental estimate—later in conjunction with the assumptions below.

**Assumption 2.** \( \gamma(\theta) \) is linear. Assumptions 1 & 2 imply

\[
Y = \beta \cdot f + \gamma_0 + \gamma_1 \cdot \theta + \epsilon
\]  

(3)

Assumption 2 states that \( \gamma(\theta) \) is linear rather than fully nonparametric. This is a relatively innocuous assumption as long as \( \gamma(\theta) \) does not exhibit highly nonlinear behavior. We estimate \( \gamma(\theta) \) nonparametrically using linear splines on sales among the consumers who see only the control ad during the experiment and on sales among exposed consumers during the before treatment.\(^{19}\) The set of individuals who see only the control ad \( (f = 0) \) includes the Control group and 7% of the Half group. Estimating equation (3) reduces the number of parameters to estimate from 51 (50 \( \gamma_\theta \) fixed effects and \( \beta \)) to three \( (\gamma_0, \gamma_1, \text{and } \beta) \). However, if you are willing to assume that \( \gamma(\theta) \) is basically constant—as Figure 6 shows is reasonable—then we avoid estimating \( \gamma_1 \) and can estimate equation (4) instead.

Figure 6 displays our estimates for the relationship \( \gamma(\theta) \) between indexed sales and ad type. The thin red line shows purchases during the two weeks of the campaign among those who only see the control ad. While Figure 6 might have some nonlinearity on the range from 0 to 4 impressions and perhaps slopes upward a bit after about 30 or 40 impressions, those slight deviations are very small relative to the ad effects we will estimate. Figure 6 also tries to establish the relative \( \gamma(\theta) \) between adtype and baseline sales using sales during various pre-treatment time windows (weeks 1-8, weeks 9-52, weeks 53-104, and the two-year total). These curves reinforce the idea that consumers’ baseline purchases vary little with \( \theta \) for \( \theta \leq 50 \). These curves are much smoother because they are based on more aggregated weeks of purchase data. In order to be conservative, Figure 6 provides the widest confidence intervals, which correspond to the two weeks of the campaign. The confidence bands for the pre-treatment aggregate sales are much narrower: one-third as wide for the aggregate two-year sales. The pre-campaign data sales reassure us that purchase behavior is relatively independent of

\(^{19}\)The splines provide local smoothing between nearby values of \( \theta \). Knots are located at \( \theta_i \in \{2, 4, 6, 8, 11, 15, 20, 30, 40, 50\} \). For robustness, we also checked the results on a larger domain of values and with less smoothing. A regression on individual dummy variables \( 1(\theta_i = \theta) \) for all \( \theta \in \{1, 2, \ldots, 100\} \) yields a qualitatively similar graph, but it has much wider confidence intervals due to the absence of smoothing and the lower density of data for \( 50 < \theta \leq 100 \).
browsing behavior over the range we are studying.

**Assumption 3.** $\gamma(\theta)$ is constant. Assumptions 1 & 3 imply

$$Y = \beta \cdot f + \gamma_0 + \epsilon$$ (4)

Assumption 3 goes further and asserts that $\gamma(\theta)$ is constant. We gain additional power and obtain smaller standard errors by using additional variation in $f$. Model (4) uses the expectation of $f$, rather than just deviations of $f$ from the Binomial expectation for the Half group and the Bernoulli expectation for the Full and Control groups. We cannot reject the null hypothesis that the $\gamma(\theta) = \gamma_0 + \gamma_1 \theta$ curve is flat.\(^\circ\) Furthermore, the confidence bands show that deviations from the average level of purchases are small: within about 10% of the overall mean, independent of the value of $\theta$. This suggests that Assumption 3 may well approximate $\gamma(\theta)$ for $\theta \leq 50$.

Assumption 3 crosses a threshold where the estimates potentially include endogenous variation from ad type $\theta$. Though Assumption 3 comes with the risk of bias, this bias can be bounded using the upper bound of the confidence interval on $\gamma(\theta)$. Using the upper bounds of confidence interval in Figure 6, we bound the endogeneity bias in the model (4) estimates to about $0.01$ out of the $0.0371$ frequency effect. Still, we make Assumption 3 cautiously because other work has documented important heterogeneity in consumer behavior along this dimension. For example, Lewis (2010) found that browsing intensity on the Yahoo! front page was highly correlated with a user’s propensity to click on an ad in the range of $0 < \theta \leq 50$, with conditional click-through rates commonly varying by a factor of two for the 30 campaigns studied.

**Assumption 4.** $h(f)$ is quadratic. Assumptions 1a), 3 & 4 imply

$$Y = \beta_1 \cdot f + \beta_2 \cdot f^2 + \gamma_0 + \epsilon$$ (5)

Assumption 4 instead allows the curve $h(f)$ to be quadratic in $f$ in order to investigate the second derivative of the frequency curve. The quadratic model allows for diminishing returns to ad frequency or ad wear-out.

Table 3 collects the ad frequency results and its columns correspond to the above equation

\(^\circ\)For example, a linear regression of $Y$ on $\theta$ produces a coefficient of 0.012 with a $p$-value of 0.75.
numbers. Column 1 therefore shows the experimental estimate, which is 3.68¢ (s.e. 1.72¢). The equation (2) estimate under Assumption 1 is 3.75¢ (s.e. 1.28¢). Equation (3) uses Assumptions 1 and 2: Equation (3) yields essentially the same estimate 3.73¢ and standard error 1.28¢. Equation (4) combines Assumptions 1 & 3 and yields the estimate is 3.71¢ with standard errors of 1.02¢. From the experimental estimates in column 1 to our preferred structural estimate in column 2, we see 41% narrower confidence intervals. Finally, column (3) combines Assumptions 1a), 3, and 4 to test the second derivative of the frequency curve. The estimated coefficient on $f^2$ turns out to be positive—implying increasing return to ad frequency—but the estimate is small in magnitude. Since we cannot reject the hypothesis that $\beta_2 = 0$ (p-value: 0.404), we find our linearity assumption on $h(f)$ to be acceptable for $0 < \theta \leq 50$. The four linear estimates are very close to each other; this reassures us that our assumptions do not introducing large biases.

4.3 Heterogeneous Treatment Effects

We examine possible heterogeneous treatment effect by consumer characteristics. These characteristics include the consumer’s proximity to the retailer, purchase recency, retailer loyalty, income, and gender. We examine characteristics for which economic theory or logic predicts heterogeneous treatment effects. At base, we reason that these characteristics—which are associated with more baseline sales at the retailer—will also make consumers more receptive to the retailer’s advertising. Recall that our preferred ad effect estimate is $0.48$ per consumer in the full treatment group and that the estimate has a t-statistic of only 2.34. When we divide the sample by consumer characteristics, the subsamples have fewer observations and this will decrease the precision of our estimates. Despite this, we can show marginally significant differences in ad effectiveness by proximity, recency, loyalty, and income. We feel our heterogeneous treatment effect estimates provide a valuable datapoint in rare such experimental studies and indicate directions for future research.

Table 4 presents our heterogeneous treatment effect estimates for consumer proximity, recency, past retail spending, income, and gender. To read the table, the estimates present the treatment effects separately by whether or not users satisfy a condition on a given consumer characteristic. For instance, the proximity condition is satisfied only when a user lives within a one mile radius of a retailer storefront. Among treated users for which we have location data, only 3.4% live within a mile of the retailer. We define recency to encompass the 23% of treated users who complete any
transaction with the retailer in the eight weeks prior to a user seeing her first ad exposure. For retailer loyalty, we set the cutoff at $1,000 in the past two years: a third of the population exceeds that amount. We then segment user income by whether a user makes more than $100,000 annually, which accounts for 52% of treated users for whom we have income data. Finally, women make up 70% of the treated population. Throughout, we use our preferred treatment effect estimator from Section 4.1. The results are similar if we choose nearby cut-off points for the continuous variables.

Table 4 suggests that the effect of advertising is concentrated in consumers who live near a store, purchase recently, spend heavily at the retailer, are wealthy, and are female. The point treatment estimates are larger in these instances for both the Full and Half group. Further, the treatment estimates for these instances in the Full group are nearly all significantly different from zero at 5% (at 10% for retailer loyalty). In the Full group, the ad effect is high for consumers within 1 mile of the retailer at $2.88 (s.e. $1.43) and low farther away at $0.49 (s.e. $0.23). An F-test of equality between these two groups is marginally rejected (p-value: 0.098). The pattern is similar for the other variables. Recent shoppers exhibit a higher effect $1.62 (s.e. $0.75) than not $0.13 (s.e. $0.14) and the two are statistically different (p-value: 0.051). Customers with high past sales have a $1.74 lift (s.e. $0.93) versus $0.16 (s.e. $0.10) in the rest, which are marginally different (p-value: 0.094). Wealthier consumers account for most of the ad effect and the difference is marginally significant (p-value: 0.090): consumers earning more than $100,000 account for a $0.81 (s.e. $0.33) effect and the rest account for $0.11 (s.e. $0.24). Women have a larger effect of $0.58 (s.e. $0.26) than men with $0.22 (s.e. $0.32), but the difference is not significantly different from zero (p-value: 0.382).

Thus, we have marginally significant evidence that ad effectiveness is heterogeneous by consumer proximity to the retailer, purchase recency, retailer loyalty, and wealth.

We find our evidence of a proximity effect to be compelling because economic theory supports the relationship and the result is replicated elsewhere. Grossman and Shapiro (1984)'s theoretical model of advertising in a spatially differentiated market supports a proximity effect. In a simple version of the model (Bagwell, 2007), consumers are distributed along a unit interval with one firm on either end. Firms can send advertising messages to all consumers that reach them with some probability. In equilibrium, both firms advertise and a firm's advertising particularly increases its sales from consumers located nearby. However, the result can also be explained within a persuasive advertising paradigm: the ads increase consumer utility sufficiently to overcome their cost of
shopping at the retailer so that consumers buy more. Alternately, the phenomenon could simply result from the complimentarity between the online ads and the store's signage. In the empirical realm, Anderson and Simester (2004) discuss suggestive evidence that closer consumers are more affected by advertising in their retail catalog ad effectiveness experiment. We also find comparable evidence of a proximity effect in unpublished experiments with two other retailers on Yahoo! These results actually inspired the development of a new product at Yahoo! Called Proximity Match, this product allows advertisers to target only those customers who live within a specified number of miles of one of a list of store locations.

We believe that our heterogenous treatment results are plausible. In all instances, the baseline sales are higher where we see a larger ad effect: this suggests that advertising succeeds where the retailer is already succeeding. Consumers who purchase less in the past and do not purchase recently may not respond to advertising because they are not in the market for apparel or prefer the retailer's competitors. Some will be tempted here to rule out informative advertising among a firm's most dedicated—and therefore best informed consumers—but we could instead interpret the advertising as an informative reminder for consumers. Wealthier customers may be more affected here because they have more money to spend and branded apparel is a luxury good. The ads predominantly focus on women’s apparel, so we expect women to be more receptive to the advertising. Our evidence vindicates marketers who target their ads based on consumer proximity, recency, spending, income, and gender.

We hope that our evidence will guide future researchers when they examine determinants of advertising effectiveness. Our study is a single data point, so we do not know if our results are general. We acknowledge that our evidence is marginally statistically significant and warrants some concern over multiple hypothesis testing. We also acknowledge that the tests in Table 4 are correlated to the extent that these characteristics are correlated; for instance, purchase recency and past expenditure are correlated. Nonetheless, we believe our evidence is valuable because ad experiments with consumer level ad exposure and sales data are rare and the availability of such consumer characteristic data is rarer still.
4.4 Miscellaneous Results

In this subsection, we collect a miscellany of results that decompose the ad effect by campaign, sales channel, shopping trips versus basket size and more. We use preferred estimator from Section 4.1 throughout. This means that the regression model includes our full set of covariates and outcome variable only includes purchases that take place after a consumer's first ad exposure.

4.4.1 Individual Campaign and Post-Campaign Impact

Table 5 considers the effect of advertising for both retailer ad campaigns individually, and after the campaigns concluded.

The first two columns of Table 5 separately examine the two campaigns in the experiment. The two weeklong 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 distinguishable statistically. 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 the lingering impact of Campaign 1, but we cannot test this hypothesis because we did not randomize treatment independently between campaigns.

The third columns of Table 5 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 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 21% and 112% 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 $0.337$ 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 increases the variance of the outcome variable. Since this increases the noise in our estimates more than the underlying signal, our results concentrate on the ad impact during the campaigns.

These longer term estimates allay somewhat the concern that the ad effect only reflects inter-temporal 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. Anderson and Simester (2004) find evidence that short run ad effects are due to this inter-temporal substitution effect among a catalog retailer’s established customers. Our estimates of the effect during and after the campaigns allay this concern somewhat. Also, Lewis and Reiley (2010)’s experiment with the same retailer found a significant impact in the week after a two-week campaign and found suggestive evidence of an impact multiple weeks after this campaign.

### 4.4.2 Sales Channel: Online Versus In-Store

Table 6 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 Lewis and Reiley (2010)’s 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 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 6.76% over the control group but in-store sales increase by only 2.97%. The proportional lift in the Half group is about the same: 1.58% for online sales and 1.70% for in-store sales.

### 4.4.3 Probability of Purchase Versus Basket Size

Marketers often wish to decompose the effect of ads on sales into increasing the probability of purchase and buying more when they do shop. Table 7 shows the experimental differences in probability of purchase and on the ‘basket size’ or purchase amount conditional on a transaction. We present the basket size results as descriptive since we cannot identify the additional consumers that arrive due to ads and therefore cannot say how the other consumer’s basket size changes as a result of the advertising. In addition, Table 8 examines the effect on ads on the number of shopping trips.

Table 7 illustrates our probability of purchase versus basket size 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% for the Full treatment group and by 0.44% 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. Table 7’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 either. Relative to a baseline mean basket size of $171, these represent percentage increases of 2.24% and 0.74%, respectively.

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.\textsuperscript{21} 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 8 shows the impact of ads on shopping trips. The Full treatment produces 0.0020 additional transactions ($p=0.013$) and the Half treatment produces 0.0011 additional transactions ($p=0.14$) 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 than for online sales. 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 Reiley (2010) 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.

\textsuperscript{21}We define a transaction to be a net positive sale or negative return.
5 Conclusion

We have conducted a large-scale experiment that exogenously varies the frequency of advertising shown to existing customers of a large retailer, making a number of improvements over previous research on the effects of advertising. We took a regularly scheduled online display advertising campaign and reallocated the budget so that one third of customers received twice as many ads as they would have otherwise (the Full treatment group), one third received a normal amount (the Half treatment group), and one third received none at all (the Control group). During the two weeks of the campaign, the Full group shows a statistically significant increase in sales of about 2.5%, while the Half group shows just under half that effect.

Even with a carefully designed experiment, millions of observations, and individually matched data on sales and advertising, the effects of advertising are extremely difficult to measure precisely; the size of the economically meaningful treatment effect is quite small compared with the variance of consumer purchases. The fact that we are able to detect statistically significant treatment effects without imposing any structural modeling assumptions is therefore quite an achievement. We accomplished this through careful experimental design and data collection. Conditioning on two years of past individual-level purchase data helped to reduce the conditional variance of purchases. We attained even larger gains in precision by judicious exclusion of irrelevant, noisy data: running control ads allowed us to exclude users who could never have been influenced by advertising, while collecting daily sales data allowed us to exclude purchases made before a customer had seen their first ad impression.

In order to be able to make more precise estimates of the effects of frequency, we then impose modeling assumptions that exploit endogenous as well as exogenous variation in ad frequency. Though such variation could in principle introduce bias, we are able to use our randomized experiment to check the validity of these assumptions and find that they seem reasonable. For consumers with more than fifty opportunities to see the ads in this campaign, we find that to exploit the endogenous variation in frequency (the amount that they browse the relevant pages on Yahoo!) would introduce considerable spurious correlation and bias of our causal estimates. However, we see no evidence of such bias for consumers with fifty or fewer impression opportunities, and exploiting their endogenous variation allows us to increase our statistical precision considerably. Restricting
attention to these consumers, we find approximately constant returns to scale for exposure frequencies between zero and fifty impressions in two weeks. With these more powerful estimates, we can also show that this advertising had a statistically significantly positive impact on profits (a hypothesis requiring much more statistical power than the hypothesis that it has a positive impact on purchases). Notably, most online display advertisers talk about optimal frequencies much lower than 50 impressions in two weeks; we have seen campaigns with maximum impressions per customer of only three per month. Our results indicate that conventional wisdom in the online advertising industry may have significantly underestimated the optimal frequency of exposure.

Several of our results are consistent with the earlier experiment of Lewis and Reiley (2010), who were unable to study exogenous effects of frequency. The overall treatment effect is similar but smaller in magnitude. We also confirm that the online advertising significantly increases offline as well as online sales, and that positive effects persist after the end of the campaign. We find that the advertising works mainly through getting individuals to make a larger number of transactions (rather than more dollars per transaction). This latter result is possible with the help of daily data that Lewis and Reiley (2010) lacked; without it, we would have missed the fact that quite a few individuals are stimulated to make a second or third visit to the store during a given two-week period.

We find some evidence of heterogeneous treatment effects across types of customers. Customers who live geographically close to a physical store tend to show larger increases in purchases due to advertising, as do customers who are wealthier, have purchased more in the past from the retailer, or have made a more recent purchase more recently from the retailer. The geographic proximity results inspired a new product at Yahoo!, called Proximity Match, whereby advertisers can target advertising only to consumers who live within several miles of one of their store locations.

Advertising is an important economic activity with many open questions. Before we can definitively answer questions about how and why advertising works, we first must be able to measure how much it affects consumer behavior, a measurement problem that turns out to be quite difficult to do with any precision. In this paper, by running a careful experiment with millions of subjects, we are able to reach the threshold of measuring statistically significant effects on revenues and profits. This despite the fact that we know our estimates are conservative; they will be attenuated by several factors, including incomplete attribution of sales to a given consumer, mismatching of consumer
retail accounts to Yahoo! accounts, and observing purchases for a time period that fails to cover all long-run effects of the advertising. We show that the optimal frequency of online display advertising appears to be much higher than that allowed by conventional wisdom in the industry. Given how difficult it is to obtain precise measurements of advertising effectiveness, we find it plausible that suboptimal firm behavior might persist. The methodology demonstrated in this paper may help us better understand advertising from the point of view of both firms and consumers.

References


, , 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.


### Table 1: Advertising Experiment Summary Statistics

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Full</th>
<th>Half</th>
<th>Control</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>1,032,204</td>
<td>1,032,074</td>
<td>1,032,299</td>
<td>0.988</td>
</tr>
<tr>
<td>Female (mean)</td>
<td>68.5%</td>
<td>68.5%</td>
<td>68.5%</td>
<td>0.794</td>
</tr>
<tr>
<td>Age (mean)</td>
<td>43.6</td>
<td>43.6</td>
<td>43.6</td>
<td>0.607</td>
</tr>
<tr>
<td>Yahoo! page views(^a) (mean)</td>
<td>122.9</td>
<td>122.2</td>
<td>121.7</td>
<td>0.135</td>
</tr>
<tr>
<td>Pre-Treatment sales (2 years, mean)</td>
<td>$857.74</td>
<td>$859.30</td>
<td>$855.54</td>
<td>0.475</td>
</tr>
<tr>
<td>Pre-Treatment sales (2 weeks, mean)</td>
<td>$19.34</td>
<td>$19.24</td>
<td>$19.10</td>
<td>0.517</td>
</tr>
</tbody>
</table>

#### Treated Subsample

**Both Campaigns**

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>Half</th>
<th>Control</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed sample</td>
<td>572,574</td>
<td>571,222</td>
<td>570,908</td>
<td>0.254</td>
</tr>
<tr>
<td>Proximity to retailer (miles, mean)</td>
<td>21.25</td>
<td>21.24</td>
<td>21.30</td>
<td>0.805</td>
</tr>
<tr>
<td>Yahoo! page views (mean)</td>
<td>268.6</td>
<td>268.5</td>
<td>266.5</td>
<td>0.108</td>
</tr>
<tr>
<td>Ad views (mean)</td>
<td>33.42</td>
<td>33.41</td>
<td>33.66</td>
<td>0.349</td>
</tr>
<tr>
<td>Ad views (median)</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Retailer ad views (mean)</td>
<td>33.42</td>
<td>16.69</td>
<td>-</td>
<td>0.801</td>
</tr>
<tr>
<td>Control ad views (mean)</td>
<td>-</td>
<td>16.72</td>
<td>33.66</td>
<td>0.165</td>
</tr>
<tr>
<td>Retailer ad click-through rate(^b)</td>
<td>0.19%</td>
<td>0.24%</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Retailer ad clicker rate(^c)</td>
<td>4.91%</td>
<td>3.39%</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

**Campaign 1**

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>Half</th>
<th>Control</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed sample</td>
<td>499,388</td>
<td>499,378</td>
<td>497,626</td>
<td>0.127</td>
</tr>
<tr>
<td>Ad views (mean)</td>
<td>20.34</td>
<td>20.33</td>
<td>20.51</td>
<td>0.072</td>
</tr>
<tr>
<td>Retailer ad views (mean)</td>
<td>20.34</td>
<td>10.16</td>
<td>-</td>
<td>0.845</td>
</tr>
<tr>
<td>Control ad views (mean)</td>
<td>-</td>
<td>10.16</td>
<td>20.51</td>
<td>0.043</td>
</tr>
<tr>
<td>Retailer ad click-through rate(^b)</td>
<td>0.168%</td>
<td>0.199%</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Retailer ad clicker rate(^c)</td>
<td>2.876%</td>
<td>1.922%</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

**Campaign 2**

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>Half</th>
<th>Control</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed sample</td>
<td>504,001</td>
<td>502,568</td>
<td>503,168</td>
<td>0.357</td>
</tr>
<tr>
<td>Ad views (mean)</td>
<td>17.81</td>
<td>17.78</td>
<td>17.90</td>
<td>0.420</td>
</tr>
<tr>
<td>Retailer ad views (mean)</td>
<td>17.81</td>
<td>8.87</td>
<td>-</td>
<td>0.501</td>
</tr>
<tr>
<td>Control ad views (mean)</td>
<td>-</td>
<td>8.91</td>
<td>17.90</td>
<td>0.377</td>
</tr>
<tr>
<td>Retailer ad click-through rate(^b)</td>
<td>0.221%</td>
<td>0.278%</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Retailer ad clicker rate(^c)</td>
<td>3.167%</td>
<td>2.308%</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.
Table 2: Effect of Advertising on Sales: Refinements in Precision
Average Treatment Effect on the Treated Estimates (Sales During Two Weeks of Experiment)

<table>
<thead>
<tr>
<th>Subset of Users&lt;sup&gt;a&lt;/sup&gt;</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales After First Ad Exposure&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Everyone</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
</tr>
<tr>
<td>Sales After First Ad Exposure&lt;sup&gt;b&lt;/sup&gt;</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Full Treatment ($)</td>
<td>0.673**</td>
<td>0.525**</td>
<td>0.559**</td>
<td>0.553**</td>
<td>0.535**</td>
<td>0.486**</td>
<td>0.477**</td>
</tr>
<tr>
<td>(0.317)</td>
<td>(0.237)</td>
<td>(0.217)</td>
<td>(0.217)</td>
<td>(0.213)</td>
<td>(0.204)</td>
<td>(0.204)</td>
<td></td>
</tr>
<tr>
<td>Half Treatment ($)</td>
<td>0.0248</td>
<td>0.189</td>
<td>0.307</td>
<td>0.307</td>
<td>0.239</td>
<td>0.241</td>
<td>0.221</td>
</tr>
<tr>
<td>(0.311)</td>
<td>(0.235)</td>
<td>(0.217)</td>
<td>(0.217)</td>
<td>(0.212)</td>
<td>(0.209)</td>
<td>(0.209)</td>
<td></td>
</tr>
<tr>
<td>Constant ($)</td>
<td>15.52***</td>
<td>15.53***</td>
<td>13.17***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.122)</td>
<td>(0.166)</td>
<td>(0.154)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Covariates**
- Demographics<sup>c</sup>
- Customer categories<sup>d</sup>
- Past sales (2 years)<sup>e</sup>
- Exposure intensity<sup>f</sup>

<table>
<thead>
<tr>
<th>Observations</th>
<th>3,096,577</th>
<th>1,714,704</th>
<th>1,714,704</th>
<th>1,714,704</th>
<th>1,714,704</th>
<th>1,714,704</th>
<th>1,714,704</th>
</tr>
</thead>
<tbody>
<tr>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.042</td>
<td>0.090</td>
<td>0.091</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. <sup>a</sup>The treated of users are those who are exposed to either the retailer or the control ad. <sup>b</sup>Sales after 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. <sup>c</sup>Two-year sales of pre-treatment—both online and in-store—at the weekly level except for aggregate sales for weeks 9 through 44 and 61 through 104. 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. <sup>d</sup>Demographic covariates include individual gender, age, state dummies as well as the user’s tenure as a Yahoo! customer. <sup>e</sup>The retailer customer category covariates include categorical variables for recency of last purchase, customer loyalty, and lifetime customer spending. <sup>f</sup>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.
Table 3: Estimates of the Effects of Advertising Frequency

<table>
<thead>
<tr>
<th>Subset of Users</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>Sales</td>
<td>Sales</td>
<td>Sales</td>
<td>Sales</td>
<td>Sales</td>
</tr>
<tr>
<td>Number of ad-views, $f_\hat{\theta} = 0.0368$</td>
<td>0.0375</td>
<td>0.0373</td>
<td>0.0371</td>
<td>0.0159</td>
<td></td>
</tr>
<tr>
<td>(0.0172)</td>
<td>(0.0128)</td>
<td>(0.0128)</td>
<td>(0.0102)</td>
<td>(0.0276)</td>
<td></td>
</tr>
<tr>
<td>ad-views squared, $f_\hat{\theta}^2$</td>
<td>0.000626</td>
<td>(0.000750)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potential Ad Views, $f_\hat{\theta} - 0.000302$</td>
<td>(0.00993)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-15.20***</td>
<td>15.19***</td>
<td>1.53***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.152)</td>
<td>(0.125)</td>
<td>(0.530)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1,395,826</td>
<td>1,395,826</td>
<td>1,395,826</td>
<td>1,395,826</td>
<td>1,395,826</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. All regressions were performed on the subsample such that $0 < \hat{\theta} \leq 50$ where $\hat{\theta}$ is the total number of the retailer's and control ad impressions seen during the two-week campaign. Subset of users are those who are exposed to either the retailer or the control ad (Treated).
Table 4: Heterogenous Treatment Effects by Consumer Attributes

<table>
<thead>
<tr>
<th>Treatment Effect Estimates Conditional on Consumer Attributes</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterogeneous Effects Variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proximity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lives within 1 mile of nearest store</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transacted in 8 weeks pre-treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loyalty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spent over $1,000 in 2 years pre-treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earns $100,000 or more</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subset of Users&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sales After First Ad Exposure&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Treatment ×</td>
<td>2.883**</td>
<td>1.624**</td>
<td>1.735*</td>
<td>0.810**</td>
<td>0.583**</td>
</tr>
<tr>
<td>Satisfies Condition</td>
<td>(1.430)</td>
<td>(0.752)</td>
<td>(0.933)</td>
<td>(0.332)</td>
<td>(0.259)</td>
</tr>
<tr>
<td>Full Treatment ×</td>
<td>0.485**</td>
<td>0.129</td>
<td>0.161</td>
<td>0.112</td>
<td>0.223</td>
</tr>
<tr>
<td>Does Not Satisfy Condition</td>
<td>(0.233)</td>
<td>(0.139)</td>
<td>(0.103)</td>
<td>(0.240)</td>
<td>(0.321)</td>
</tr>
<tr>
<td>Half Treatment ×</td>
<td>1.523</td>
<td>0.804</td>
<td>0.481</td>
<td>0.383</td>
<td>0.347</td>
</tr>
<tr>
<td>Satisfies Condition</td>
<td>(1.528)</td>
<td>(0.776)</td>
<td>(0.959)</td>
<td>(0.339)</td>
<td>(0.267)</td>
</tr>
<tr>
<td>Half Treatment ×</td>
<td>0.340</td>
<td>0.0485</td>
<td>0.156</td>
<td>0.0463</td>
<td>-0.0501</td>
</tr>
<tr>
<td>Does Not Satisfy Condition</td>
<td>(0.240)</td>
<td>(0.138)</td>
<td>(0.102)</td>
<td>(0.242)</td>
<td>(0.315)</td>
</tr>
<tr>
<td>Covariates: Full Set&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Non-missing cond. data&lt;sup&gt;d&lt;/sup&gt;</td>
<td>77.1%</td>
<td>100%</td>
<td>100%</td>
<td>98.2%</td>
<td>99.6%</td>
</tr>
<tr>
<td>Observations</td>
<td>1,714,704</td>
<td>1,714,704</td>
<td>1,714,704</td>
<td>1,714,704</td>
<td>1,714,704</td>
</tr>
<tr>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.103</td>
<td>0.103</td>
<td>0.103</td>
<td>0.103</td>
<td>0.103</td>
</tr>
</tbody>
</table>

Estimates are for the average treatment effect during the two weeks of the experiment conditional on the relevant consumer characteristics. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. <sup>a</sup>The treated of users are those who are exposed to either the retailer or the control ad. <sup>b</sup>Sales after 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. <sup>c</sup>Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table X for details). We also include indicator variables for the given condition. <sup>d</sup>The data for some variables (distance, income, gender) is incomplete. We include all observations however as these improve the estimates for the covariates.
Table 5: Effects of the Advertising During and After the Campaign
Average Treatment Effect on the Treated Estimates (Sales)

<table>
<thead>
<tr>
<th>Timeframe</th>
<th>Campaign 1</th>
<th>Campaign 2</th>
<th>During &amp; After Campaigns Total (4 weeks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset of Users(^a)</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
</tr>
<tr>
<td>Sales After First Ad Exposure(^b)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Full Treatment ($)</td>
<td>0.116</td>
<td>0.382**</td>
<td>0.525*</td>
</tr>
<tr>
<td></td>
<td>(0.144)</td>
<td>(0.153)</td>
<td>(0.309)</td>
</tr>
<tr>
<td>Half Treatment ($)</td>
<td>0.059</td>
<td>0.156</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>(0.141)</td>
<td>(0.155)</td>
<td>(0.312)</td>
</tr>
<tr>
<td>Covariates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics(^c)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Customer categories(^d)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Past sales (2 years)(^e)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Exposure intensity(^f)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Observations</td>
<td>1,496,392</td>
<td>1,509,737</td>
<td>1,714,704</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.058</td>
<td>0.056</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses. *** \(p<0.01\), ** \(p<0.05\), * \(p<0.1\). \(^a\)The treated of users are those who are exposed to either the retailer or the control ad. \(^b\)Sales after 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. \(^c\)Two-year sales of pre-treatment—both online and in-store—at the weekly level except for aggregate sales for weeks 9 through 44 and 61 through 104. 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. \(^d\)Demographic covariates include individual gender, age, state dummies as well as the user’s tenure as a Yahoo! customer. \(^e\)The retailer customer category covariates include categorical variables for recency of last purchase, customer loyalty, and lifetime customer spending. \(^f\)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.
Table 6: Effects of the Advertising, Online versus Offline Average Treatment Effect on the Treated Estimates (In-Store and Online Sales During Two Weeks of Experiment)

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>All Sales</th>
<th>In-Store Sales</th>
<th>Online Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset of Users&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
</tr>
<tr>
<td>Sales After First Ad Exposure&lt;sup&gt;b&lt;/sup&gt;</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>Full Treatment ($)</strong></td>
<td>0.477**</td>
<td>0.323*</td>
<td>0.154**</td>
</tr>
<tr>
<td></td>
<td>(0.204)</td>
<td>(0.172)</td>
<td>(0.0779)</td>
</tr>
<tr>
<td><strong>Half Treatment ($)</strong></td>
<td>0.221</td>
<td>0.185</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>(0.209)</td>
<td>(0.176)</td>
<td>(0.081)</td>
</tr>
</tbody>
</table>

**Covariates**

| Demographics<sup>c</sup> | x | x | x |
| Customer categories<sup>d</sup> | x | x | x |
| Past sales (2 years)<sup>e</sup> | x | x | x |
| Exposure intensity<sup>f</sup> | x | x | x |

Observations 1,714,704 1,714,704 1,714,704

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. <sup>a</sup>The treated of users are those who are exposed to either the retailer or the control ad. <sup>b</sup>Sales after 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. <sup>c</sup>Two-year sales of pre-treatment—both online and in-store—at the weekly level except for aggregate sales for weeks 9 through 44 and 61 through 104. 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. <sup>d</sup>Demographic covariates include individual gender, age, state dummies as well as the user’s tenure as a Yahoo! customer. <sup>e</sup>The retailer customer category covariates include categorical variables for recency of last purchase, customer loyalty, and lifetime customer spending. <sup>f</sup>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.
Table 7: Effects of the Advertising, Probability of Purchase versus Basket Size Average Treatment Effect on the Treated Estimates (During Two Weeks of Experiment)

<table>
<thead>
<tr>
<th>Subset of Users&lt;sup&gt;a&lt;/sup&gt;</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales After First Ad Exposure&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Treated</td>
<td>Treated</td>
<td>Transacted</td>
</tr>
<tr>
<td>Full Treatment ($)</td>
<td>0.477**</td>
<td>0.000426</td>
<td>3.822</td>
</tr>
<tr>
<td></td>
<td>(0.204)</td>
<td>(0.000461)</td>
<td>(2.391)</td>
</tr>
<tr>
<td>Half Treatment ($)</td>
<td>0.221</td>
<td>0.000474</td>
<td>1.365</td>
</tr>
<tr>
<td></td>
<td>(0.209)</td>
<td>(0.000462)</td>
<td>(2.438)</td>
</tr>
</tbody>
</table>

Covariates

- Demographics<sup>c</sup>  
- Customer categories<sup>d</sup>  
- Past sales (2 years)<sup>e</sup>  
- Exposure intensity<sup>f</sup>

Observations 1,714,704 1,714,704 132,568  
$R^2$ 0.091 0.147 0.107

Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.  
<sup>a</sup>The treated of users are those who are exposed to either the retailer or the control ad.  
<sup>b</sup>Sales after 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.  
<sup>c</sup>Two-year sales of pre-treatment—both online and in-store—at the weekly level except for aggregate sales for weeks 9 through 44 and 61 through 104. 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.  
<sup>d</sup>Demographic covariates include individual gender, age, state dummies as well as the user’s tenure as a Yahoo! customer.  
<sup>e</sup>The retailer customer category covariates include categorical variables for recency of last purchase, customer loyalty, and lifetime customer spending.  
<sup>f</sup>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.
Table 8: Effects of the Advertising: Purchase Size versus Probability of Purchase
Average Treatment Effect on the Treated Estimates (Probability of Transaction During Two Weeks of Experiment)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In-Store + Online</td>
<td># of Days with a Transaction</td>
<td># of Days with a Transaction</td>
<td>In-Store + Online</td>
<td># of Days with a Positive Net Transaction</td>
<td># of Days with a Positive Net Transaction</td>
</tr>
<tr>
<td>Dependent Variable</td>
<td>All Sales</td>
<td>In-Store Sales</td>
<td>Online Sales</td>
<td>All Sales</td>
<td>In-Store Sales</td>
<td>Online Sales</td>
</tr>
<tr>
<td>Subset of Users</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
</tr>
<tr>
<td>Sales After First Ad Exposure</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Full Treatment ($)</td>
<td>0.00196**</td>
<td>0.00129*</td>
<td>0.000662***</td>
<td>0.00191***</td>
<td>0.00124**</td>
<td>0.000662***</td>
</tr>
<tr>
<td></td>
<td>(0.000795)</td>
<td>(0.000698)</td>
<td>(0.000212)</td>
<td>(0.000721)</td>
<td>(0.000620)</td>
<td>(0.000212)</td>
</tr>
<tr>
<td>Half Treatment ($)</td>
<td>0.001162</td>
<td>0.000956</td>
<td>0.000205</td>
<td>0.001205*</td>
<td>0.000999</td>
<td>0.000205</td>
</tr>
<tr>
<td></td>
<td>0.001162</td>
<td>0.000956</td>
<td>0.000205</td>
<td>0.001205*</td>
<td>0.000999</td>
<td>0.000205</td>
</tr>
</tbody>
</table>

**Covariates**
- Demographics
  - x
- Customer categories
  - x
- Past sales (2 years)
  - x
- Exposure intensity
  - x

Observations | 1,714,704 | 1,714,704 | 1,714,704 | 1,714,704 | 1,714,704 | 1,714,704 |
R² | 0.174 | 0.171 | 0.090 | 0.147 | 0.143 | 0.090 |

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. aThe treated of users are those who are exposed to either the retailer or the control ad. bSales after 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. cTwo-year sales of pre-treatment—both online and in-store—at the weekly level except for aggregate sales for weeks 9 through 44 and 61 through 104. 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. dDemographic covariates include individual gender, age, state dummies as well as the user’s tenure as a Yahoo! customer. eThe retailer customer category covariates include categorical variables for recency of last purchase, customer loyalty, and lifetime customer spending. fThe 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: Histogram of Customers’ Distance to the Nearest Store by Treatment Assignment

Figure 2: Histogram of Total Ad Exposures (Both Retailer and Control Ads)
Figure 3: Histogram of Average Weekly Purchases During the Two Years Before the Experiment

Figure 4: Histogram of Customer Purchases in the Two Weeks Before the Experiment (Conditional on Transacting)
Figure 5: Histogram of Customer Purchases During the Two-Week Ad Experiment (Conditional on Transacting)

Figure 6: Indexed Sales Given a Customer’s Experimental Ad Exposures ($\theta$)
Figure 7: Marginal Impact of Ad Frequency Estimates $\beta_\theta$ by Ad Type ($\theta$)

Impact of Display Ad Frequency on Sales

- Average Effect of Frequency
- 95% Confidence Intervals

* Estimates obtained by regressing sales on ad views, interacted with dummies for the discrete number of potential ad views, $\theta_{\text{ad views}}$. 