Location, Location, Location: Repetition and Proximity Increase Advertising Effectiveness

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

March 5, 2015

Abstract

Yahoo! Labs partnered with a nationwide retailer to study the effects of online display advertising on both online and in-store purchases. We measure the impact of frequency of advertising exposure using a simple randomized experiment: 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. We find strong benefits to repeated exposures among users who see up to 50 ads, with revenues increasing approximately linearly at a rate of 4¢ per exposure. We find especially high ad effectiveness for the retailer's best customers and those who live closest to its brick-and-mortar locations; these findings are consistent with advertising in a Hotelling model of differentiated firms.

Keywords: advertising effectiveness, field experiments, digital advertising, ad frequency, targeting

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

1 Introduction

From 2007 to 2011, Yahoo! Labs conducted nineteen large-scale controlled experiments to measure the effects of online advertising on retail sales. The experiment reported in this paper represents the best of these, incorporating wisdom accumulated during other experiments and breaking new ground on scale. We examine two consecutive weeklong ad campaigns that target three million of the retailer's existing customers, matching subsequent retail purchases with advertising exposure at the individual level. These retail image ads featured attractive photos of featured items, but no prices, detailed product information, or calls to action. Our experimental estimates suggest the retailer ads increased sales by 3.6% and that the campaigns were profitable. On the question of ad frequency, we experimentally vary the intensity of treatment and find less "wear out" than many expect: constant returns to advertising for as many as 50 ad exposures in two weeks. We also show that the retailer ads generate the largest sales lift among the retailer's best customers and those who live near a store, documenting that the retailer's benefits to advertising increase with the consumers' proximity to the store.

Yahoo Labs! emphasized the use of controlled experiments in order to avoid endogeneity between advertising and consumer behavior, having documented serious bias from methods commonly used in the advertising industry to analyze observational data. Observational studies, typically comparing endogenously exposed versus unexposed users, may overstate ad effectiveness by orders of magnitude (Lewis, Rao and Reiley, 2011) or even have the wrong sign relative to experimental estimates (Lewis and Reiley, 2014). Advertisers' advertising choices can induce bias, for instance, by targeting customers who are more likely to purchase or by targeting times like Christmas when customers purchase more. Lewis, Rao and Reiley (2011) show that online consumer choices can also induce bias (what they term 'activity bias') because consumers' activity is correlated across websites without that correlation necessarily being causal.

Controlled experiments remain rare in the advertising industry. This dearth of experiments seems surprising given the dollars at stake: advertising represents between 1% and 2% of global GDP (Bughin and Spittaels, 2012), and U.S. online display advertising revenues alone reached $7.9 billion in 2013, excluding mobile (IAB, 2014). We believe that the potential pitfalls of endogeneity and selection bias in observational studies are not well appreciated by industry analysts, despite a report commissioned by the Interactive Advertising Bureau (Lavrakas, 2010) affirming experiments as the gold standard for measurement.

Retailers account for a large share of online display ads (8% of total impressions, according to (comScore, 2013)), but multi-channel retailers face challenges in measuring the in-store sales impact of their online advertising. To help solve this problem, Yahoo! Labs partnered with five retailers and with third-party data-matching services in order to link consumer data from Yahoo! on ad views and a retailer's in-store and offline sales data. This proved to be a nontrivial task: a number of the experiments encountered technical problems, rendering data unusable or statistically underpowered.

The Yahoo! Labs experiments demonstrate that even large ad experiments have low statistical power. Lewis and Rao (2013) critically examine these Yahoo! Labs advertising experiments. They point out a 'signal-to-noise' problem: the mean impact of advertising is plausibly a small fraction of the variance in sales. Such ad experiments have low statistical power and may require millions of observations—across consumers, campaigns, or time—to detect a profitable impact.

The experiment in this paper stands out for its large size and spending as well as its superior experimental design and execution. First, our experiment includes 3 million eligible users—double the size of the previous experiment with the same retailer (Lewis and Reiley, 2014). Second, our experiment includes control ads that boost the precision of our estimates. The control ads also allow us to separate the impact of ad frequency from the user's eligibility to see ads based on her endogenous browsing intensity. Third, the experiment features exceptional data on consumers, including proximity to the retailer as well as past consumer spending at the retailer, which boost precision and identify heterogeneous treatment effects.

Our innovations improve statistical power, allowing us to tackle new research questions. We use the control ads to identify the counterfactual exposures in the control group so that we can filter out both unexposed users and pre-exposure sales in our experimental difference. Removing these two sources of noise increases our precision by 31% without inducing bias. We improve a further 5% by including numerous covariates in our linear regression in order to reduce the residual variance. At this stage, we can show that the campaign statistically significantly increases sales. To further establish that the campaign is profitable, we impose and test restrictions on the functional form of the effect of ad frequency.

Other advertising experiments deal with the statistical-power problem in different ways. (Lodish et al., 1995) pioneered split-cable television experiments with a panel of 3,000 households with advertising treatment matched to household-level purchases of consumer packaged goods (Lodish et al., 1995). With merely thousands of consumers, these studies lacked statistical power, with the majority of 600 such tests failing to demonstrate statistically significant effects (Lodish et al., 1995; Hu et al., 2007). Sahni (2013) gains power by restricting attention to consumers who have already expressed purchase intent in his study of the effects of restaurant display ads on lead generation on a restaurant search website. Blake et al. (2013) similarly examine consumers whose searches imply purchase intent, varying treatment over time or across cities rather than across individual consumers. In studying the effects of retail image advertising, we are interested in the effects of image ads on consumers who have not yet expressed purchase intent; our efforts to increase power are therefore crucial.

We designed our experiment to examine the shape of the consumer ad frequency response curve. The experiment includes a ‘Full’ treatment group that is exposed to the retailer’s ads and a ‘Control’ group that is exposed to unrelated, control ads. We also included a ‘Half’ treatment group that sees a retailer or control ad with equal probability at each ad viewing opportunity. Our experimental estimates suggest an average sales lift of $0.477 on the Full treatment group and $0.221 on the Half treatment group, which represent a 3.6% and a 1.7% increase over the Control group’s sales. We find constant returns to ad frequency for users who see up to 50 total retailer and control ads. We measure the marginal value of an additional ad to be 3.68¢ for these users.

Most previous research on the returns to ad frequency relies on observational data or lab experiments (see Pechmann and Stewart, 1988, for a survey), while a few recent papers use experimental variation in the field. Lewis (2010) examines the frequency curves of online display ads' impact on clicks and conversions using natural experiments reaching tens of millions of Yahoo! users in a single day. He finds heterogeneous frequency effects among campaigns: one-third demonstrate constant returns for over 20 ad exposures in a single day, though the rest exhibit some degree of wear-out. Sahni (2013) experimentally varies treatment intensity on a restaurant search website and finds decreasing returns to online display ads after only three exposures in a session. The results on returns to scale might depend on differences in purchase intent: Sahni (2013) finds decreasing returns to frequency of ads delivered during related product searches, while Lewis (2010) and our study find more constant returns for ads delivered during casual browsing.

Finally, we consider how advertising effectiveness may vary with customer proximity, in a Hotelling model of differentiated firms. Motivated by the model of Grossman and Shapiro (1984), we predict that a firm's advertising most affects consumers who are 'near' the firm or are already inclined toward a firm's product. We test this implication using variation in the consumer's geographic proximity to the retailer's stores, as well as in the consumer's proximity in taste for the retailer's product offering. We examine measures of RFM (recency, frequency and monetary value) to indicate consumers' revealed preferences for the firm, as well as consumer income as a measure of segmentation. We confirm our prediction that proximity mediates ad effectiveness in all five proximity measures. Our geographic proximity results accord with those of Molitor et al. (2013) who find that mobile coupon redemption rates increase as consumers get closer to the store. Our RFM results contrast with experimental studies in catalog advertising (Simester et al., 2009) and search advertising (Blake et al., 2013) that find advertising is less effective on the advertiser's best customers. We speculate that our results differ due to the difference in customer intent associated with display media and because our retailer advertises infrequently, which reduces cumulative wear-out.

The rest of this paper is organized as follows. The next section describes our experimental design and data collection. The third section provides descriptive statistics and an illustrative power calculation. The fourth section presents our measurements of the causal effects of the advertising, focusing particularly on the effects of frequency and heterogeneous treatment effects by consumer proximity. 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 in spring 2010. In each week, the advertisements featured a different line of branded clothing. The experimental subjects were randomly assigned to treatment groups that remained constant for both weeks. A confidentiality agreement prevents us from naming either the retailer or the featured product lines.

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

The retailer ads are primarily image advertising and are co-branded with apparel firms. The ads display the store brand, the brand of the featured product line, and photographs of the advertised clothing line on attractive models. The ads do not include any price or promotion information.

[1] These campaigns were purchased on an impression, not click, basis. This avoids distortions induced by ad servers changing delivery patterns in the treatment and control groups due to the retailer and control ads having different click-through rates. Such distortions render the control ads unsuitable for tagging counterfactual exposures in the control group.

Figure 1 presents an illustrative example with the retailer Target and the designer Missoni—neither of which is involved in the experiment. The creative content of each ad impression is dynamic and involves slideshow-style transitions between still photographs and text. Campaign 1 includes three different ads in an equal-probability rotation: men's apparel, women's apparel, and women's shoes. Campaign 2 advertises a product line from a different manufacturer and features women's apparel. Consequently, we acknowledge that our ad frequency results refer to repeated ad exposures from the same retailer but reflect significant creative rotation. The control ads advertise Yahoo! Search and are depicted in Figure 2.

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

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

3 Data

The data section contains two 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. Appendix A.1 details the source of our data and some key variables.

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 and 55% of users were exposed to an experimental ad. Throughout Table 1, F-tests of equality of means of the variables by treatment group are not rejected, which is consistent with a valid experimental randomization. In each treatment group 68.5% of customers are female, the average age is 43.6 years, and customers viewed an average of 245 web pages on Yahoo! during the campaign. Customers spent an average of \$19.23 at the retailer during the two weeks prior to the experiment and \$857.53 in the two years beforehand. Figure 3 shows that the distribution of average weekly sales over the two years prior to the experiment are essentially identical across all three treatment groups. On average, treated consumers live 21 miles from the nearest retailer location, but half live within 6.6 miles.

In Table 1, we see that the experiment delivers advertisements evenly across treatment groups. For the remainder of this section, our descriptive statistics exclude the 45% of users who were not exposed. Figure 4 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 right, 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.

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).

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 the meta-analysis by Lewis and Rao (2013) of the Yahoo! Labs ad experiments with sales data.

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 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

$$\begin{aligned} H_0 : \Delta sales &= \$0 \\ H_a : \Delta sales &= \$0.51, \end{aligned}$$

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. In Section 4.1, we present methods that reduce the standard deviation of sales to \$111. The expected *t*-statistic is given by

$$E[T] = \frac{\hat{\delta} - \delta_0}{SE\left(\hat{\delta}\right)} = \frac{0.51 - 0}{\frac{111\sqrt{2}}{\sqrt{570{,}000}}} = 2.45$$

for the average treatment effect estimator $\hat{\delta}$.[2] Using the 10% two-sided critical value of $t^{\star} = 1.645$, the power of the test is given by

$$Pr(T > t^{\star}|E[T] = 2.45) = 79\%$$

The power calculation elucidates both the limits of our experimental results and the importance of improving the precision of our treatment effect estimates.

The statistical power for our main effect is 79%. Without the econometric methods described in Section 4.1, our power would be much lower at 49%. Under the above assumptions, the statistical power for the Half group—or for distinguishing the Full and Half groups—is 34%. If we seek to detect longer run ad effects, this compounds the statistical power problem. For instance, if the hypothesized \$0.51 lift occurs against the background noise of more than just the two weeks of sales during the campaign, our statistical power will be reduced (Lewis et al., 2014). We emphasize that the above calculations 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. In Section 4.2, we develop a model of ad frequency that further increases power and therefore managerial confidence under some assumptions.

[2] Let $\hat{\mu}$ denote the sample average and σ^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

$$Var\left(\hat{\delta}\right) = Var\left(\hat{\mu}_T - \hat{\mu}_C\right) = \frac{\sigma_T^2}{N_T} + \frac{\sigma_C^2}{N_C} = \frac{2\sigma^2}{N}$$

where we assume that $N_T = N_C$ and $\sigma_T^2 = \sigma_C^2$, which is a good approximation here.

4 Results

The results section is divided into three subsections, which we briefly preview. Section 4.1 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. Our preferred experimental estimates suggest that consumers who were exposed to the retailer's ad saw their average sales increase by $\$0.477$ (3.6%) in the Full treatment group and $\$0.221$ (1.7%) in the Half treatment group. Section 4.2 analyzes the role of ad frequency on ad effectiveness and finds constant returns of 3.7¢ per ad up to 50 ads. To show this, we restrict the functional form of the ad frequency effects and show that these restrictions are supported by the data. Section 4.3 examines heterogeneous treatment effects by consumer proximity to the retailer. We find large and statistically significant ad effects within the subpopulation who live within one mile of a retailer, transact within eight weeks, represent the top third of purchase frequency at the retailer, spend more than $1,000 at the retailer in the previous two years, or earn more than $100,000. In Appendix A.2, we separate the effect of advertising by campaign, sales channel, and shopping trips. We find that the majority of the total treatment effect is attributable to the in-store rather than online sales channel and estimate a 1.8% increase in shopping trips in the Full group. Appendix A.2 also shows that including sales after the campaign increases the ad effect estimates, which allays the concern that the in-campaign estimates merely reflect intertemporal substitution of purchases.

4.1 Overall Campaign Impact

Table 2 presents regression estimates of the average effect of treatment on the treated (TOT) for the impact of advertising on consumer purchases during the two-week experiment. In particular, Table 2 highlights the various methods for increasing the precision of the estimates without inducing bias. These methods include pruning components of the outcome data that cannot 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 TOT estimate. The indirect TOT 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).[3] The indirect TOT estimator relies on the fact that outcomes among untreated subjects have an expected experimental difference of zero. However, the difference among untreated subjects adds noise to the estimator. The indirect TOT estimator yields a $0.67 average sales lift (s.e. $0.32) in the Full treatment group and an average lift of $0.03 (s.e. $0.31) in the Half group. Whereas Lewis and Reiley (2014) estimate the TOT indirectly out of necessity, this experiment employs control ads to identify the counterfactual treated sample in the Control group.

Table 2's column (2) presents the direct TOT regression estimate on the treated sample. The treated sample are those users who see any of the experiment's retailer or control ads during the two weeks of the campaign. The direct TOT estimator increases precision by pruning the untreated subsample which contributes only noise to the estimator. The regression 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 TOT requires control ads to identify the treated subgroup in the Control, 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 cannot influence sales 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) in the Full group and

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

\$0.31 (s.e.: \$0.22) in the Half group.

In column 4 of Table 2, we increase the precision of our estimates by adding covariates to the TOT regression. The covariates in the regression improve precision by reducing the unexplained variance in the dependent variable.[4] Specifically, these covariates include: demographics, retailer-defined customer segments, consumer sales history, and browsing intensity. The demographic covariates include indicator variables for gender, year of age, and state of residence as well as a scalar variable for the time since the consumer signed up with Yahoo!. The retailer-defined RFM customer segments include Recency of last purchase, Frequency of past purchases, and Monetary Value (total lifetime spending at the retailer). We include two years of individual-level past sales data and pre-treatment sales during the campaign, separately for online and offline sales. The browsing intensity covariates are 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.

Including covariates improves precision by 5% (columns 3–4) across Full and Half groups, while pruning irrelevant data improves precision by 31% (columns 1–3) on average.[5] However, covariates may be inexpensive to include and serve to validate the experimental randomization. Control ads are expensive but facilitate data pruning and thereby improve precision five times more than covariates.

Our preferred experimental estimator—in column (4) of Table 2—measures a $\$0.477$ (s.e.: $\$0.204$) increase in average sales from the ads in the Full group and a $\$0.221$ (s.e. $\$0.209$) increase in the Half group. Though all the estimates in Table 2 are unbiased, we prefer the estimates in column (4) as they are the most precise. The point estimates with covariates are more conservative because they are smaller than those in column (3). The point estimates in column (4) likely fall because they account for differences like the slightly lower sales in the Control group in the two weeks before treatment (see Table 1). The preferred Full treatment effect is statistically significant at

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

[5] The historical sales and customer category covariates account for nearly all of the 5% improvement. The demographic and exposure intensity covariates provide almost no precision improvement.

the 5% level (p-value: 0.020) and represents a 3.6% sales lift over the Control group. The Half treatment effect is not significantly different from zero (p-value: 0.289), and the joint F-test is marginally significant (p-value: 0.065). The point estimates indicate that doubling the advertising approximately doubles the effect on purchases, but the precision of these estimates is too low 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.

To make decisions about advertising, managers not only want to establish a revenue lift, but also calculate return on investment. 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 $\$273,000 \pm 229,000$, while the Half campaign increased purchases by $\$126,000 \pm 234,000$, using 95% confidence intervals. Compared with costs of about \$88,000 for the Full campaign and \$44,000 for the Half campaign, these indicate incremental revenues of around three times the cost. We assume a gross margin of 50% for the retailer's sales.[6] Our point estimates indicate a rate of return of 51% on the advertising dollars spent, but with a 95% confidence interval of [-101%, 204%].

Our short-run sales effect estimates are likely conservative as several factors attenuate the result. These factors include: 1) incomplete attribution of sales to a given consumer, 2) unseen ad impressions due to ad-blocking or ads below the fold, 3) mismatching of consumer retail accounts to Yahoo! accounts, 4) logged-in exposures viewed by other household members, and 5) observing purchases for a time period that fails to cover all long-run effects of the advertising. Though short-run effects could outpace the long-run effects due to intertemporal substitution as in Simester et al. (2009), our estimates in Section 4.2 and Appendix A.2.1 that include sales for the two weeks post-treatment suggest a positive long-run effect.

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

4.2 Effects of Ad Frequency

In this section, we quantify the marginal value of an additional ad impression. Above, we learned that doubling the number of ad impressions to the treatment group increases the estimated treatment effect from $\$0.221$ to $\$0.477$ per person. However, the confidence intervals on these point estimates are sufficiently wide that we cannot 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 measure the ad frequency response curve. We find that the frequency curve is linear for users who can see up to 50 ads. For these users, we estimate the marginal impact of an ad to be $3.71¢$, which is eight times the cost per ad of $0.46¢$. Beyond understanding ad frequency response curves, we wish to judiciously model ad response in order to improve managerial confidence in advertising decisions. Including sales both during and after the campaign, we use our frequency model to calculate a 138% return on investment (ROI) for the retailer's ad campaign.

To measure ad frequency response curves, we want to distinguish between the advertiser's decision to purchase more ads and the user's decision to visit more web pages on Yahoo!. Both choices increase the chance that a user sees more ads, but the advertiser can only impact the former. Towards this, we introduce the two 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 θ—to mean her potential ad exposures given by the sum of her control and retailer ad exposures. The user's browsing intensity endogenously determines her ad type.

The data contain three sources of variation in ad frequency, which are depicted in Figure 5. First, the three treatment groups provide exogenous variation in frequency. 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, $f \sim Binomial(0.5, \theta)$. The variation in f within the Half group is less than 5% as large as the variation between treatment groups: the number of retailer ad exposures clusters around $\frac{1}{2}\theta$ for users with $\theta \geq 10$ as Figure 5 shows. Third, user-level variation in ad type θ—due to browsing intensity—provides variation in f. Unlike the other two sources of variation in f, the variation due to θ 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 81% of treated subjects. As Figure 4 illustrates, data on users who see more than 50 impressions rapidly becomes sparse, with less than 2% seeing more than 200 impressions. Our frequency effect estimates 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. We show estimates for $\theta > 50$ in Section 4.2.2.

4.2.1 Frequency Model Estimates

As a starting point we estimate the model

$$Y = \beta \cdot f + \gamma_0 + \gamma_1 \cdot \theta + \epsilon. \tag{1}$$

This simple linear model allows sales (Y) to depend on ad type through γ_1 and tells us the marginal value of an additional retail impression through β. This simple model fits the data well. While we allow for the possibility of wear-in or wear-out in Section 4.2.2, we are unable to reject the constant returns to ad frequency for $0 < \theta \leq 50$. Throughout, 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.

Column (1) of Table 3 shows that the marginal impact of a retailer ad is 3.73¢ (s.e. 1.28¢) while the marginal impact of ad type is -0.03¢ (s.e. 0.99¢). The results tell us that ad type seems to play a secondary role for $0 < \theta \leq 50$. We say this because the coefficient on ad type is small and insignificant. Also, we cannot reject the null hypothesis that the $\gamma(\theta) = \gamma_0 + \gamma_1\theta$ curve is flat.[7] Figure 6 nonparametrically estimates the relationship between ad type θ and indexed sales among exposed users. Figure 6 shows that sales both during the experiment (for users with $f = 0$) and

[7] A linear regression of sales on θ for $f = 0$ produces a coefficient of 0.4¢ with a p-value of 0.75.

before the experiment appear unrelated to θ.

Given that ad type θ is not too important, we assume away baseline purchase heterogeneity by imposing $\gamma_1 = 0$ and estimate

$$Y = \beta \cdot f + \gamma_0 + \epsilon. \tag{2}$$

Model (2) uses the variation in observed value 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.

Imposing $\gamma_1 = 0$ may be a poor assumption in general because β then includes endogenous variation from ad type θ but appears to work here for $0 < \theta \leq 50$. The assumption fails when we include users who do not see ads ($\theta = 0$) since sales increase discontinuously from $\theta = 0$ to $\theta = 1$. For users with $\theta > 50$, either $\gamma_1 \neq 0$ or frequency is nonlinear since the experimental difference estimate is negative but insignificant for that subpopulation.

Assuming $\gamma_1 = 0$ improves our precision with minimal bias. We estimate model (2) in column (2) of Table 3 and see that the marginal effect of an ad is almost unchanged at $3.71¢$ but that the standard error falls by 25% to $1.02¢$. In column (3), we estimate the marginal effect of an ad on total sales both during the two-week campaign and the two weeks afterwards. Our four-week estimate suggests that the average effect of an ad is $6.12¢$ (s.e. $1.58¢$) for the 80% of consumers with $0 < \theta \leq 50$.

Model (2) delivers statistically significant evidence that the campaign is profitable—thereby boosting managerial confidence. Using the column (3) estimates, we conclude that the campaign generates profits of \$628,000 with 95% confidence intervals between [\$310,000, \$946,000] and return on investment of [18%, 258%], centered at 138%. To arrive at this figure, we multiply our four-week estimate by the total impressions delivered to users with $0 < \theta \leq 50$ and the retailer's gross margin of 50% (see footnote 6). Because our point estimate for the lift among the $\theta > 50$ group is negative but insignificant, we assume no lift among that subgroup because we reject that the ads reduced sales. We arrive at profits by subtracting the total cost of the campaign (\$132,000), including the cost of users with $\theta > 50$ to be conservative. Without the frequency model, neither our estimates in Section 4.1 nor the earlier Lewis and Reiley (2014) experiment show significant evidence of profitability.

In our view, measuring ad effectiveness necessitates some compromise between experimental evidence and modeling assumptions. Experiments can deliver unbiased estimates of the effects of advertising. However, Lewis and Rao (2013) show that this setting is so low-powered that experiments require impractical sample sizes in the millions or billions to optimize ROI. On the other hand, modeling ad effectiveness in the absence of experiments may create biases, and even small biases are important given the effect size of advertising. We believe that the way forward is to use experimental evidence to formulate and verify models of ad effectiveness. The experimental and modeling approaches can thereby complement each other to explore the challenging task of optimizing ad spend.

In our experiment, we find approximately constant returns to advertising out to 50 impressions over the course of two weeks, a high figure. This result is difficult to compare to the existing experimental literature on ad frequency. Lewis (2010)'s meta-study estimates frequency curves for up to 50 impressions on a single day of advertising on Yahoo!'s front page using binomial variation in frequency alone. He finds that a minority of campaigns—a third—demonstrate constant returns for clicks up to 20 exposures and finds some campaigns with constant returns up to 50 impressions. Nevertheless, the majority of the campaigns with decreasing returns showed at most moderate levels of wear-out. Note that Lewis (2010) rejects the $\gamma_1 = 0$ assumption in that setting: propensity to click is correlated with the user's ad type. Sahni (2013) finds decreasing returns to clicking after only three ad exposures in a browsing session on a restaurant search website. However, both these studies use the immediate outcome of clicks rather than sales and consider a short time horizon with high exposure intensity. Further, unlike those in Sahni (2013), the ad exposures in our experiment on Yahoo! are unrelated to purchase intent. Future research could examine the generalizability of these conclusions. In particular, we wonder what role creative rotation plays in ad wear out: here, the experimental ads arise from two separate co-branded campaigns and the first campaign includes three kinds of ad creatives.

4.2.2 Frequency Model Robustness

We first wish to re-examine our constant returns to advertising assumption. We do so by estimating equation $Y = \beta_1 \cdot f + \beta_2 \cdot f^2 + \gamma_0 + \epsilon$ in column (4) of Table 3. Our estimated coefficient on f^2 is 0.063¢ (s.e. 0.075¢) suggesting increasing rather than decreasing returns to ad frequency. However, the estimate for the curvature in ad frequency is small and insignificant. Still, we cannot exclude the possibility of wear-in or wear-out here because some amount of curvature is consistent with the wide confidence intervals of our linear frequency estimates in columns (1) and (2). Rather, our estimates suggest that a linear model of frequency well approximates the returns to frequency for this campaign.

Next, we consider a more general model of the form $Y = \beta_\theta \cdot f + \gamma_\theta + \epsilon$ that allows the returns to frequency f to interact with ad type θ. This model makes transparent use of the experiment to estimate the returns to frequency. To illustrate, consider the users with ad type $\theta = 10$. We fit a line with slope β_{10} to the sales among the control group users with $f = 0$, the full treatment users with $f = 10$, and the half group users who are binomially distributed around $f = 5$. Thus, β_θ can be interpreted as the marginal value of a retailer ad f for a given θ. This 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 θ. Figure 7 collects the 50 slope parameter estimates β_θ for $0 < \theta \leq 50$. We see that these parameters are noisy but typically positive and our prefered estimate of 3.71¢ lies within the 95% confidence intervals of nearly all the β_θ estimates. Though lower θ have more observations, the confidence intervals are tighter for higher θ because they contain greater variation in f that increases leverage under the linearity assumption. The estimates exhibit no clear trend that would indicate increasing or decreasing returns to advertising. Finally, column (5) of Table 3 provides an impression-weighted average estimate of these 50 slope parameters of 3.68¢ (s.e. 1.72¢) which is within 1% of the more restrictive models in columns (1) and (2). We see wide confidence intervals under this more general model, which illustrates the power problem for this and even more general models.

Figure 8 shows non-parametric estimates for frequency effects using the model $Y = \beta_f \cdot f + \gamma_\theta + \epsilon$. Unlike the above model that varies frequency effect by θ, this model regresses sales on 50 dummy variables corresponding to each level of f. The individual $\beta_{f'}$ estimates correspond to the difference in average sales between users who see f' retail ads and those who see $f = 0$ for users with $0 < \theta \leq 50$. Figure 8 also includes a line with the slope of our marginal frequency estimate 3.7¢ that intersects with almost all the confidence intervals of the $\beta_{f'}$ estimates. Once again, we see that the linear model well approximates a more general model.

Next, 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 experimental specification from Section 4.1, but aggregating the Half and Full groups into a single treatment group. We obtain a noisy average treatment effect estimate of -5.8¢ (s.e. 48.3¢) 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 89¢ by 88, which yields 1¢ as an upper bound on the average treatment effect per ad exposure for users with $\theta > 50$. The marginal frequency effect estimate is thus much lower for $\theta > 50$ than our estimate of 3.71¢ for $0 < \theta \leq 50$. Hence, non-experimental models that fail to observe or condition on θ risk large bias. As we mention above, the low marginal ad effect for high θ could either be due to diminishing marginal returns to frequency or due to lower ad effectiveness for high θ users. Since f and θ are highly correlated, we cannot distinguish between the two explanations. Note that the cut-off of $\theta = 50$ is approximate: the results are similar if we set the cut-off at $\theta = 40$ or $\theta = 60$. Hence, 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. Such a frequency cap is much larger than the weekly caps of about 3 online display ads that we have seen in industry.

4.3 Advertising Effectiveness in Differentiated Markets

As Goldfarb (2013) notes, unparalleled targeting capability distinguishes online advertising and promises to improve ad effectiveness. To improve the marketing manager's future ad targeting, we empirically test for differences in ad effectiveness between consumer segments. To guide us, we consider a model of competitive advertising in differentiated markets by Grossman and Shapiro (1984) and extended by Soberman (2004). We test a key implication for consumer behavior in these models using multiple measures of proximity. In each case, we find significant evidence that consumers nearest to the differentiated firm respond most to its advertising.

In the Bagwell (2007) distillation of the Grossman and Shapiro (1984) model, two firms compete from opposite ends of a unit interval over uniformly distributed consumers in the familiar Hotelling model of horizontally differentiated firms. In the model, consumers are unaware of either firm and therefore have no baseline sales, though the ads could instead serve as a reminder. Firm i chooses advertising intensity ϕ_i that informs individual consumers with probability ϕ_i uniformly along the unit interval. Under standard assumptions, the symmetric firms each advertise $\phi > 0$ in equilibrium. Each firm then captures a fraction $\phi(1-\phi)$ of consumers who see its advertising but not its rival's ad. Among the fraction ϕ^2 of consumers who see both firms' ads, the firms split the market down the middle. Though Grossman and Shapiro (1984) model the firm's ad choices, the model's key implication for consumer behaviour is that consumers near the differentiated firm respond more to its advertising than consumers farther away by the fraction $\frac{\phi^2}{2} > 0$ of consumers. We therefore seek to reject the following null hypothesis for the experimental sales differences $\Delta sales$ by 'near' and 'far':

$$H_0 : \Delta sales_{near} \quad < \quad \Delta sales_{far}$$

We test the above hypothesis along the two most common dimensions of differentiation between firms: differentiation by consumer geography and by consumer taste. As our measure of geographic proximity, we use the crow-flies distance from the consumer's address to the retailer's nearest brick and mortar location. We consider four measures of consumer taste or preference proximity to the retailer. The first three are measures of preference proximity. These are the 'RFM' variables used in direct marketing: the *recency*, *frequency*, and *monetary value* (total amount) of consumer's past purchases. As our fourth measure, we use consumer income to split users according to the upscale retailer's target segment.

The experimental literature exhibits mixed results for the effect of ads by taste and geographic proximity. In the direct-marketing industry, RFM is thought to be a key determinant of a mailing's success (Hughes, 2000). In an experiment, Simester et al. (2009) finds instead that more catalog mailings induce the firm's best customers to shift their purchases over time but the long run effect is smaller among these customers. Blake et al. (2013) find that paid search ads are less effective among the advertiser's most recent and frequent customers whereas Raj (1982) finds the opposite for a consumer-packaged-goods advertiser's best customers with TV advertising. Sahni et al. (2014) finds that e-mail offers targeted at high monetary value users are effective. Ad experiments examining ad effects by location are rare. In a mobile advertising experiment, Molitor et al. (2013) find that coupon redemption rates are higher among consumers who are geographically close to the advertiser. Our study presents evidence of heterogeneous ad effectiveness for both measures of taste proximity and geographic proximity.

Table 4 presents our heterogeneous treatment estimates for ad effectiveness in a differentiated market. To read the table, the estimates present the treatment effects separately by whether the consumers are 'near' the firm or 'far' as defined by the variable in each column. For instance, in the first column, the near condition is satisfied only when a consumer lives within a one mile radius of a 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. We use the retailer-generated frequency categories to define the frequency cut-off as those customers in the top 3 categories, which comprise a third of users. For monetary value, we set the spending cutoff at $1,000 over 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. Throughout, we use our preferred experimental treatment effect estimator from Section 4.1. As we show in Appendix A.3, the results are similar if we choose nearby threshold values for the variables.

Table 4 suggests that the effect of advertising is concentrated in consumers who live near a store, purchased recently, spent heavily at the retailer, and are wealthy. The treatment estimates are larger in these instances for both the Full and Half group. Further, the treatment estimates for the 'near' member of the Full group are almost all significantly different from zero at 5% (10% for monetary value). 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 for those farther away at $0.49 (s.e. $0.23). We also find comparable evidence of a proximity effect in unpublished experiments with two other retailers on Yahoo!. The pattern is similar for the other variables. Recent shoppers exhibit a higher effect of $1.62 (s.e. $0.75) than those who are not $0.13 (s.e. $0.14). Frequent shoppers also have greater sales lift of $1.19 (s.e. $0.57) rather than $0.11 (s.e. $0.11). Customers with high monetary value have a $1.74 lift (s.e. $0.93) versus $0.16 (s.e. $0.10) in the rest. Wealthier consumers account for most of the ad effect: consumers earning more than $100,000 account for an $0.81 (s.e. $0.33) effect, and the rest account for $0.11 (s.e. $0.24).

In Table 4, we formally test the null hypothesis that the sales lift is larger for consumers who are farther from the retailer. We use an F-test of inequality for the coefficients on the sales lift among the 'near' consumers versus the 'far' consumers. In the Full treatment group, we are able to reject the null hypothesis at the 5% significance level for all measures: distance (p-value = 0.049), recency (0.025), frequency (0.031), monetary value (0.047), and income (0.045). In the Half treatment group, we are unable to reject the null hypothesis at the 10% significance level: distance (p-value = 0.222), recency (0.169), frequency (0.301), monetary value (0.368), and income (0.210). We then test the combined hypotheses for both treatment groups and can reject the null hypothesis at the 10% significance level: distance (p-value = 0.064), recency (0.037), frequency (0.039), monetary value (0.051), and income (0.059). Hence, we verify a testable prediction of the theory of advertising in a differentiated market for all five of our proximity measures.

While the Grossman and Shapiro (1984) model of informative advertising in differentiated markets provides an appealing explanation, other explanations are also consistent with our results. Some readers may object to the 'informative' nature of the story: existing customers by definition already know about the advertiser. We caution against such a literal interpretation and instead propose that ads could serve as a reminder or suggestion to consumers. In any model of advertising in differentiated markets, the proximity result requires the realistic assumption that consumers near the retailer can increase their purchases at the retailer. Grossman and Shapiro (1984) achieve this by assuming consumers are ignorant prior to advertising. With full information, we can construct persuasive advertising models with this feature by allowing the consumer's fixed valuation for the good to vary along a second dimension or by stealing business from a competing firm that is co-located with the advertiser.[8] However, if advertisers have too little margin to increase sales among nearby consumers—for instance, by oversupplying advertising or advertising to consumers whose 'closets are stuffed'—the proximity result could reverse. The proximity result can alternately be explained by some other kind of complementarity between baseline sales and ad effectiveness. The geographical proximity result can also be due to complementarity between advertising and the store's own signage.

We hope that this evidence on increasing advertising effectiveness with proximity will guide future researchers when they examine determinants of advertising effectiveness. We look forward to future replications that may shed light on the generalizability of this study. We caution that greater short run impact among close consumers does not necessarily imply greater long run impact. 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, the RFM variables 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

[8] Apparel retailers face multiple competitors in their target segment. These competitors usually co-locate at malls.

still. Finally, our results have implications for the firm's equilibrium advertising and pricing decisions in a differentiated market, which Iyer et al. (2005) discuss in their theoretical paper on targeted advertising.

5 Conclusion

Department store merchant and advertising pioneer John Wanamaker once remarked: "Half the money I spend on advertising is wasted; the trouble is I don't know which half." This paper tackles a classic endogeneity problem in marketing—the effect of advertising on demand—and resolves it using a clean experiment. Wanamaker's lament remains a problem—our modern-day retailer's experiment cannot reliably detect the effects of a profitable ad at normal exposure intensity (i.e., the Half treatment). Specifically, an experiment with over a million treated users would only reject a zero ad effect at the 10% significance level about 34% of the time. Lest Mr. Wanamaker despair, we show that doubling the treatment intensity allows him to detect an effective campaign 79% of the time. Moreover, Mr. Wanamaker can further improve these odds by testing and imposing restrictions on the effects of ad frequency. Finally, we empirically establish the theoretical prediction that proximity makes advertising more effective in differentiated markets: a result that would help Mr. Wanamaker find his wasted half.

Our paper also challenges some academics' heuristics in the advertising domain. One heuristic is that an experiment featuring millions of subjects should more than suffice to test theories of advertising. Nonetheless, even well-designed, large-scale experiments can be underpowered to reject even basic hypotheses like a non-zero sales effect of advertising. Additionally, our study challenges a second heuristic: profit-motivated advertisers advertise at optimal levels. As Lewis and Rao (2013) elaborate, hundreds of millions of subjects may be required to experimentally evaluate a hypothesis positing a 10% profit on ad spending. Hence, unbiased ad experiments can be both expensive and only weakly informative about profitability. More broadly, where learning is expensive we expect marketing managers to advertise at suboptimal levels—the large ROI estimates in this paper and the very small ROI estimates by Blake et al. (2013) exemplify marketing managers who may have spent too little and too much on their respective advertising strategies.

Unbiased field experiments measuring the effectiveness of advertising can contribute both to the science of consumer choice and to the advertising decisions of managers. For instance, this paper's results inspired a new product at Yahoo! called Proximity Match, which enables advertisers to target consumers who live near their stores. We anticipate that continued technological advances will reduce the cost of such studies. We look forward to future research that explores the robustness of our findings and answers other questions about how advertising influences consumer choice.

References

Bagwell, K., "The economic analysis of advertising," *Handbook of industrial organization*, 2007, *3*, 1701–1844.

Blake, Thomas, Chris Nosko, and Steven Tadelis, "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment," *NBER Working Paper*, 2013, pp. 1–26.

Bughin, Jacques and Steven Spittaels, "Advertising as an economic-growth engine," Technical Report, McKinsey & Company, http://www.mckinsey.com/locations/Belux/ /media/Belux/FinalAdvertising.ashx March 2012.

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

Goldfarb, Avi, "What is different about online advertising?," *Review of Industrial Organization*, 2013, pp. 1–15.

Grossman, Gene M. and Carl Shapiro, "Informative Advertising with Differentiated Products," *The Review of Economic Studies*, 1984, *51* (1), pp. 63–81.

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

Hughes, Arthur Middleton, "Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-based Marketing Program," 2000.

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

Iyer, Ganesh, David Soberman, and J. Miguel Villas-Boas, "The Targeting of Advertising," *Marketing Science*, 2005, *24* (3), pp. 461–476.

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

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

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

Lewis, Randall A and Justin M Rao, "On the Near Impossibility of Measuring the Returns to Advertising," *Working paper*, 2013.

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

Lewis, Randall, Justin M Rao, and David H Reiley, "Measuring the Effects of Advertising: The Digital Frontier," Technical Report, National Bureau of Economic Research 2014.

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

Molitor, Dominik, Philipp Reichhart, Martin Spann, and Anindya Ghose, "Measuring the Effectiveness of Location-Based Advertising: A Randomized Field Experiment," 2013.

Pechmann, Cornelia and David W Stewart, "Advertising repetition: A critical review of wearin and wearout," *Current issues and research in advertising*, 1988, *11* (1-2), 285–329.

Raj, S. P., "The Effects of Advertising on High and Low Loyalty Consumer Segments," *Journal of Consumer Research*, 1982, *9* (1), pp. 77–89.

Sahni, Navdeep, "Advertising Spillovers: Field Experimental Evidence and Implications for Returns from Advertising," January 2013. Preliminary draft.

_ , P.K. Chintagunta, and Dan Zou, "Effects of Targeted Promotions: Evidence from Field Experiments," 2014.

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

Soberman, David A., "Additional Learning and Implications on the Role of Informative Advertising," *Management Science*, 2004, *50* (12), 1744–1750.

6 Figures & Tables

Table 1: Advertising Experiment Summary Statistics

	Treatment Group			
	Full	Half	Control	*p*-value
Sample size	1,032,204	1,032,074	1,032,299	0.988
Female (mean)	68.5%	68.5%	68.5%	0.794
Age (mean)	43.6	43.6	43.6	0.607
Yahoo! page views[a] (mean)	245.8	244.4	243.5	0.132[d]
Pre-Treatment sales (2 years, mean)	\$857.74	\$859.30	\$855.54	0.475
Pre-Treatment sales (2 weeks, mean)	\$19.34	\$19.24	\$19.10	0.517
Treated Subsample				
Both Campaigns				
Exposed sample	572,574	571,222	570,908	0.254
Proximity to retailer (miles, mean)	21.25	21.24	21.30	0.805
Yahoo! page views (mean)	412.2	411.5	410.1	0.108
Ad views (mean)	33.42	33.41	33.66	0.164
Ad views (median)	15	15	15	
Retailer ad views (mean)	33.42	16.69	-	0.801
Control ad views (mean)	-	16.72	33.66	0.165
Retailer ad click-through rate[b]	0.19%	0.24%	-	
Retailer ad clicker rate[c]	4.91%	3.39%	-	
Campaign 1				
Exposed sample	499,388	499,378	497,626	0.127
Ad views (mean)	20.34	20.33	20.51	0.072
Retailer ad views (mean)	20.34	10.16	-	0.845
Control ad views (mean)	-	10.16	20.51	0.043
Retailer ad click-through rate[b]	0.168%	0.199%	-	
Retailer ad clicker rate[c]	2.876%	1.922%	-	
Campaign 2				
Exposed sample	504,001	502,568	503,168	0.357
Ad views (mean)	17.81	17.78	17.90	0.420
Retailer ad views (mean)	17.81	8.87	-	0.501
Control ad views (mean)	-	8.91	17.90	0.377
Retailer ad click-through rate[b]	0.221%	0.278%	-	
Retailer ad clicker rate[c]	3.167%	2.308%	-	

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

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

	(1)	(2)	(3)	(4)
Subset of Users[a]	Everyone	Treated	Treated	Treated
Sales After First Ad Exposure[b]			x	x
Full Treatment ($)	0.673**	0.525**	0.559**	0.477**
	(0.317)	(0.237)	(0.217)	(0.204)
Half Treatment ($)	0.0248	0.189	0.307	0.221
	(0.311)	(0.235)	(0.217)	(0.209)
Constant ($)	15.52***	15.53***	13.17***	
	(0.122)	(0.166)	(0.154)	
Covariates				
Demographics[c]				x
Customer categories[d]				x
Past sales (2 years)[e]				x
Exposure intensity[f]				x
Observations	3,096,577	1,714,704	1,714,704	1,714,704
R^2	0.000	0.000	0.000	0.091

Average effect of Treatment on the Treated estimates. Dependent variable is sales during the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Demographic covariates include individual gender, age, state dummies as well as the user's tenure as a Yahoo! customer. [d]Two-year sales of pre-treatment—both online and in-store—at the weekly level except for aggregate sales for weeks 9–44 and 61–104. 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. [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 3: Estimates of the Effects of Advertising Frequency Up to 50 Potential Exposures

	(1)	(2)	(3)	(4)	(5)
Dependent Variable	Sales	Sales	4-Week Sales	Sales	Sales
Subset of Users[a]	$0 < \theta \leq 50$	$0 < \theta \leq 50$	$0 < \theta \leq 50$	$0 < \theta \leq 50$	$0 < \theta \leq 50$
β: Number of Ad Views, f	0.0373***	0.0371***	0.0612***	0.0159	$\bar{\beta}$ [b]= 0.0368***
	(0.0128)	(0.0102)	(0.158)	(0.0276)	(0.0172)
β: Ad Views Squared, f^2				0.000626	
				(0.000750)	
γ_θ , γ_1 Potential Ad Views, θ	-0.000302				50 $\hat{\gamma}_\theta$'s
	(0.00993)				
Constant	15.20***	15.19***	32.24***	1.53***	-
	(0.152)	(0.125)	(0.193)	(0.530)	
Observations	1,395,826	1,395,826	1,395,826	1,395,826	1,395,826
R-squared	0.000	0.000	0.000	0.000	0.000

Robust standard errors in parentheses. *** *p*<0.01, ** *p*<0.05, * *p*<0.1. [a] All regressions were performed on the treated subsample such that $0 < \theta \leq 50$ where θ is the total number of the retailer's and control ad impressions seen during the two-week campaign. [b] $\bar{\beta} = \sum w_\theta \hat{\beta}_\theta$ and $Var\left(\bar{\beta}\right) = w_\theta^2 Var\left(\hat{\beta}_\theta\right)$ where $w_\theta = \frac{\sum_{i \in \theta} f_i}{\sum_i f_i}$.

Table 4: Heterogenous Treatment Effects by Consumer Proximity to the Retailer

	(1)	(2)	(3)	(4)	(5)
Heterogeneous Effects Variable	**Distance**	**Recency**	**Frequency**	**Monetary Value**	**Income**
Condition	*Lives within 1 mile of nearest store*	*Transacted in 8 weeks pre-treatment*	*Belongs to top 3 retailer frequency category*[e]	*Spent over $1,000 in 2 years pre-treatment*	*Earns $100,000 or more*
Subset of Users[a]	Treated	Treated	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x	x	x
‘Near’ in Full Treatment	2.883**	1.624**	1.189**	1.735*	0.810**
	(1.430)	(0.752)	(0.567)	(0.933)	(0.332)
‘Far’ in Full Treatment	0.485**	0.129	0.111	0.161	0.112
	(0.233)	(0.139)	(0.107)	(0.103)	(0.240)
‘Near’ in Half Treatment	1.523	0.804	0.425	0.481	0.383
	(1.528)	(0.776)	(0.580)	(0.959)	(0.339)
‘Far’ in Half Treatment	0.340	0.0485	0.117	0.156	0.0463
	(0.240)	(0.138)	(0.104)	(0.102)	(0.242)
Covariates: Full Set[c]	x	x	x	x	x
$H_0 : \Delta sales_{near} < \Delta sales_{far}$					
Full: F-test (p-value)	0.049	0.025	0.031	0.047	0.045
Half: F-test (p-value)	0.222	0.169	0.301	0.368	0.210
Full & Half: F-test (p-value)	0.064	0.037	0.039	0.051	0.059
% Non-missing cond. data[d]	77.1%	100%	100%	100%	98.2%
Observations	1,714,704	1,714,704	1,714,704	1,714,704	1,714,704
R^2	0.103	0.103	0.103	0.103	0.103

Average effect of Treatment on the Treated estimates conditional on the relevant consumer characteristics. Dependent variable is sales during the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details). We also include indicator variables for the given condition. [d]The data for some variables (distance, income, gender) is incomplete. We include all observations however as these improve the estimates for the covariates. [e]The frequency variables are derived from the retailer's loyalty categories: 1) Fanatic; 2) Frequent; 3) Occasional; 4) Trial; 5) No History.

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

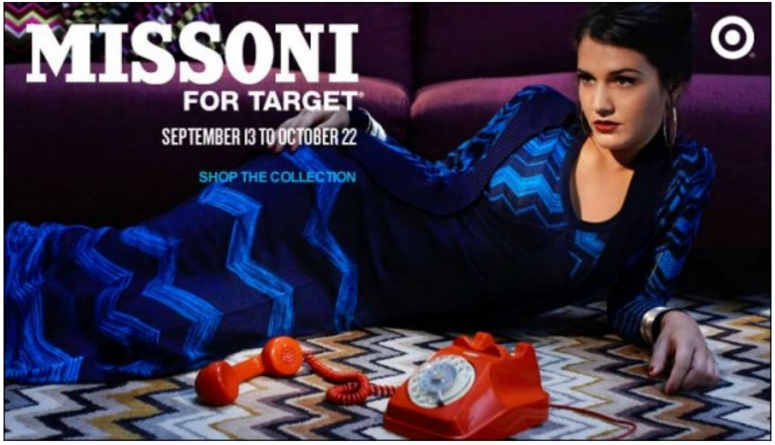


Figure 2: Control Ad: Promoting Yahoo! Search

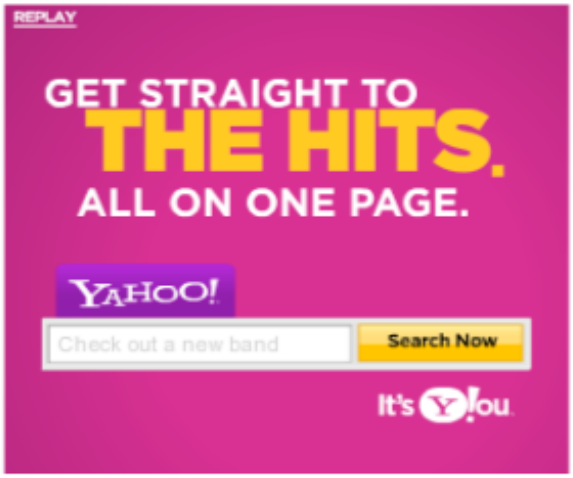


Figure 3: Histogram of Average Weekly Purchases During the Two Years Before the Experiment

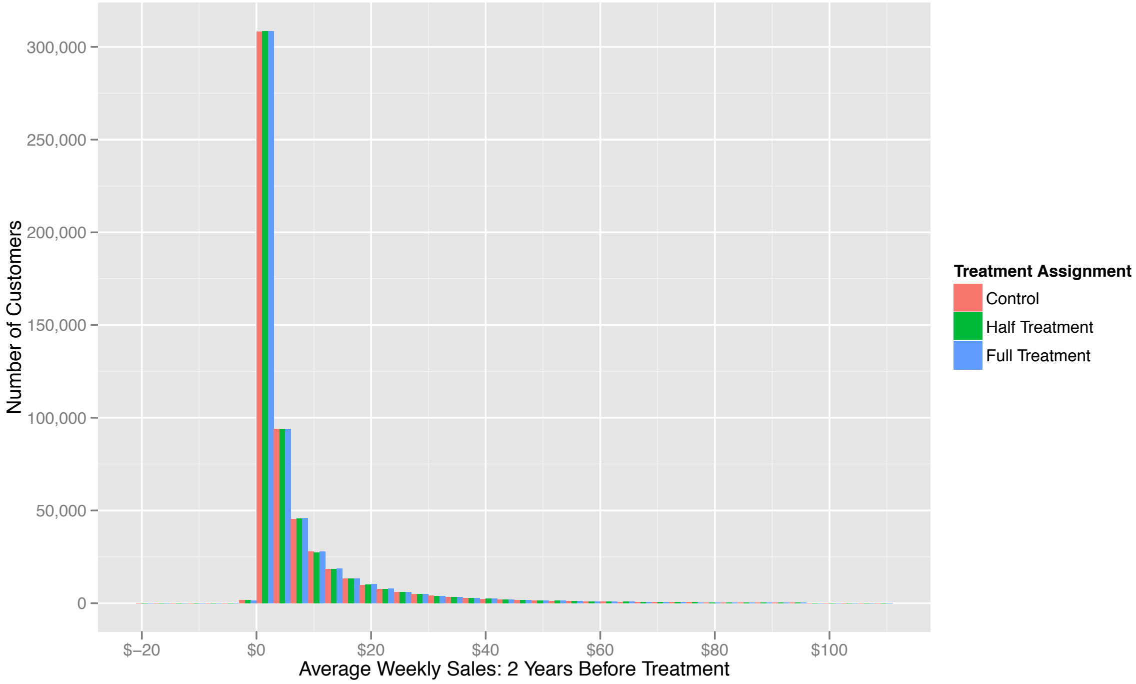


Figure 4: Histogram of Total Ad Exposures (Both Retailer and Control Ads)

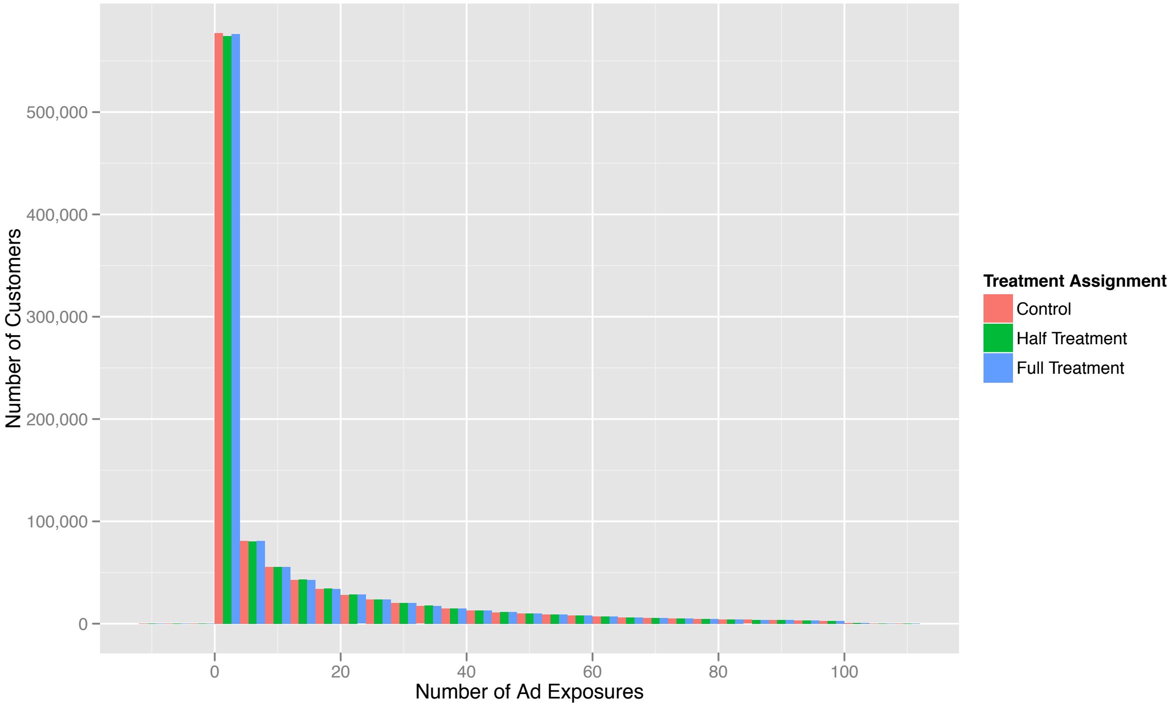


Figure 5: Joint Distribution of Retailer Ad Views (f) and Ad Type (θ)

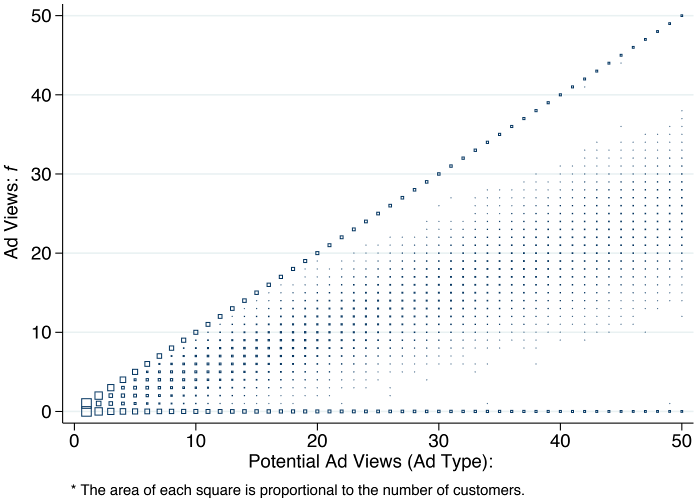


Figure 6: Indexed Sales Given a Customer's Experimental Ad Exposures (θ)

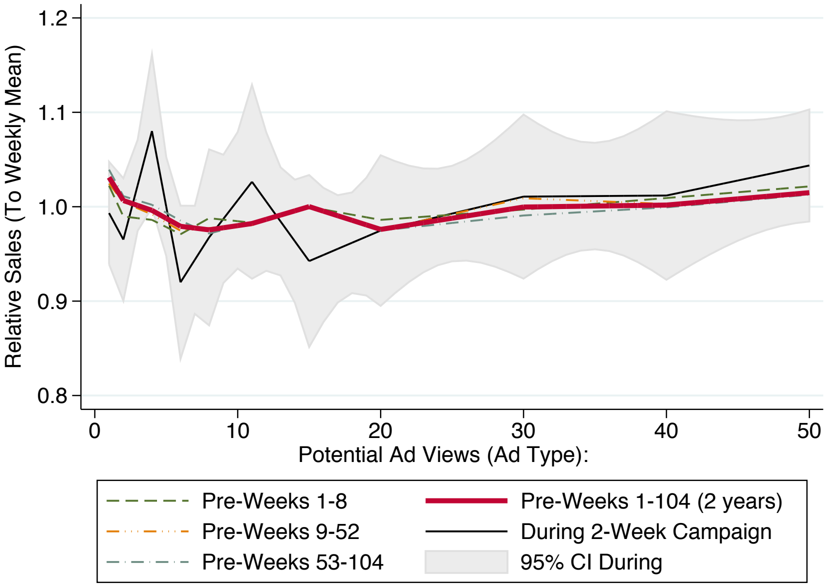
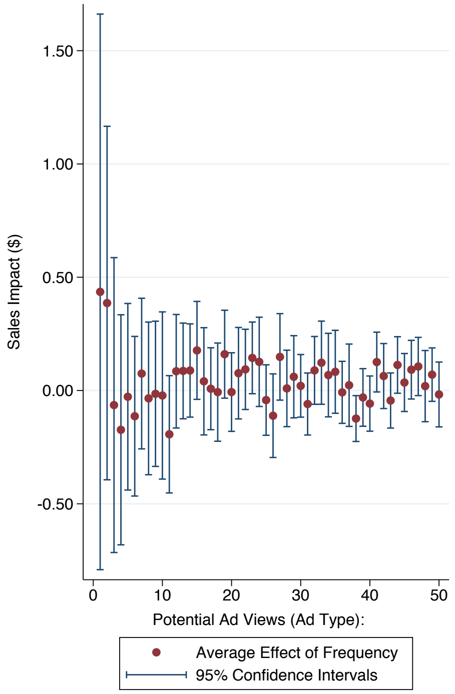
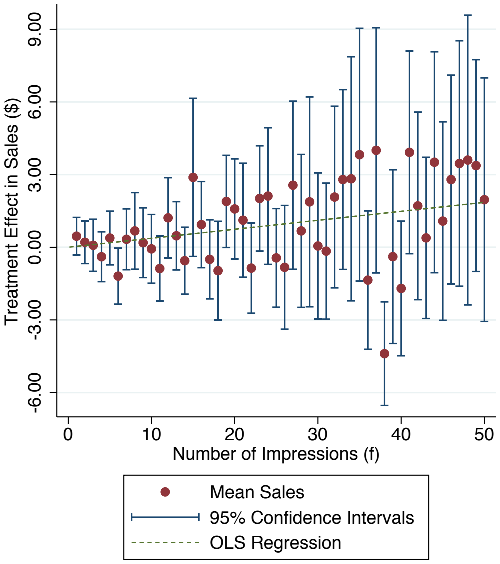


Figure 7: Marginal Impact of Ad Frequency Estimates β_θ by Ad Type (θ)



* Estimates obtained by regressing sales on ad views interacted with dummies for potential ad views (ad type),   <=50.

Figure 8: Frequency Response Curve: Treatment Effect for Each Value of f



* Estimates obtained by regressing sales on dummies, imposing the restriction of no heterogeneity by ad type for ad type   <=50.

A Online Appendix

A.1 Data Remarks

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

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

We improve on the statistical precision of (Lewis and Reiley, 2014) by collecting both more granular sales data and sales data over a longer period of time. First, we obtain daily rather than weekly transactions during the ad campaigns. Daily transaction data allow us to discard purchases that take place prior to a customer's first ad exposure. Since pre-exposure transactions could not be influenced by the advertising, including such transactions in our treatment effect estimates only adds noise. This strategy avoids sample-selection problems, because the control ads identify the corresponding pre-treatment sales in the control group.[9] Second, we obtain consumer purchase data for the two years prior to the experiment.[10] We use the purchase history as covariates in our

[9] 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.

[10] 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.

TOT 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. Each nine-digit zip code denotes a fraction of a city block, providing 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. Figure 9 shows the distribution of the consumer's proximity to the nearest retailer, which is nearly identical across treatment groups. For each customer, we compute the 'crow-flies' distance to the nearest store using the haversine formula.[11]

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

Due to an unanticipated problem in randomly assigning treatment to multiple matched consumers, we exclude almost 170,000 users from our analysis. In particular, the third-party data collection firm joined 3,443,624 unique retailer identifiers with 3,263,875 unique Yahoo! identifiers; as a result, tens of thousands of Yahoo! identifiers were matched with multiple retail identifiers. The third party performed the experimental randomization on the retailer identifiers, but provided Yahoo! with only separate lists of Yahoo! identifiers for each treatment group to book the campaigns. Some multiple matched Yahoo! users were therefore accidentally booked into multiple treatment groups, which contaminated the experiment. To avoid this contamination, we discard all the Yahoo! identifiers who are matched with multiple retailer identifiers.[12] Fortunately, the

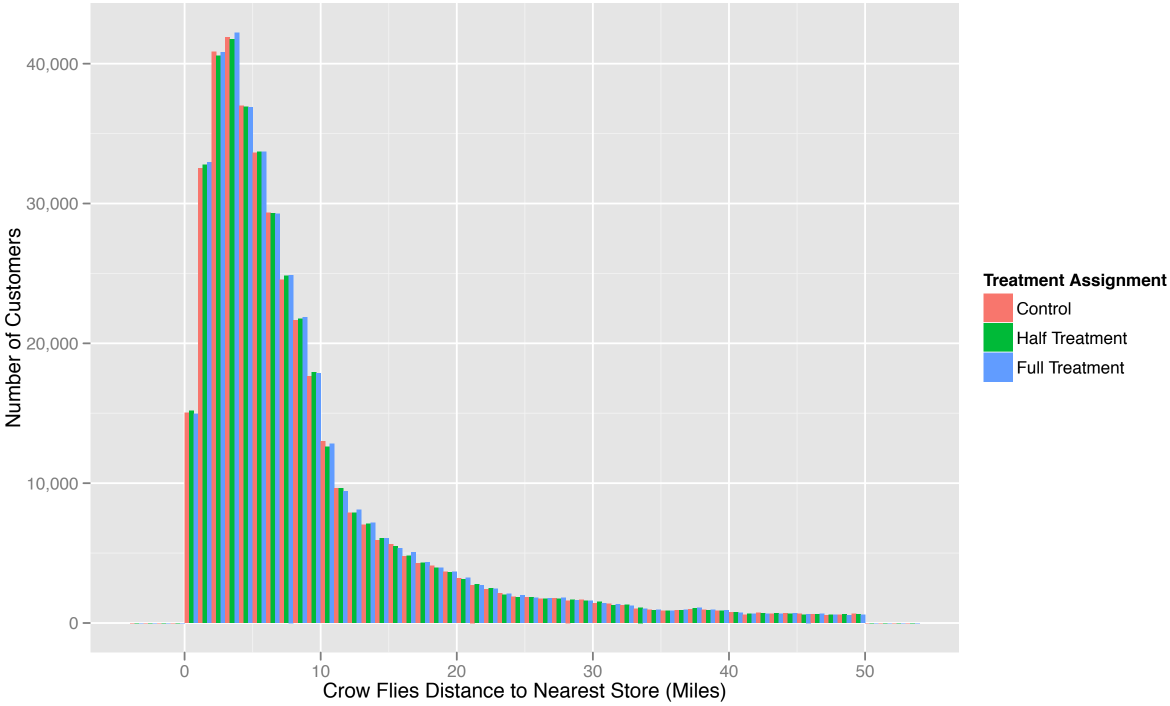
[11] The haversine formula calculates the distance between two pairs of latitude and longitude coordinates while correcting for the spherical curvature of the planet.

[12] 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.

treatment-group assignment is random, so the omitted consumers do not bias the experimental estimates. The remaining 3,096,577 uniquely matched Yahoo! users represent our experimental subjects. We acknowledge that our results only reflect 93% of exposed users.

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

Figure 9: Histogram of Customers' Distance to the Nearest Store by Treatment Assignment



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

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

A.2.1 Individual Campaign and Post-Campaign Impact

Table 5 considers the effect of advertising for both retailer ad campaigns individually and includes sales 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 statistically distinguishable. Only the Full group during Campaign 2 demonstrates a statistically significant ad effect (p-value=0.012). Some of Campaign 2's success may be due to the lingering impact of Campaign 1, but we cannot test this hypothesis because we did not randomize treatment independently between campaigns.

The third column of Table 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 (the only post-campaign data available) and the total sales impact during and after the campaign. The point estimates for the Full and Half treatment groups indicate that the total campaign impact is respectively 10% and 64% larger when we include sales after the campaign. The total ad impact is marginally statistically significant for the Full group: \$0.525 for the Full group (p-value=0.089) and \$0.363 for the Half group (p-value=0.245). Note that the standard errors are higher than in our two-week estimates in Table 2, because the additional sales data increase the variance of the outcome variable. Since this increases the noise in our estimates more than the underlying signal, we treat these positive point estimates as suggestive of lingering effects; Section 4.2 presents more powerful evidence using the ad frequency model, showing that the marginal effect of an impression on sales increases by 65% when including sales from the two weeks following the campaigns.

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

A.2.2 Sales Channel: Online Versus In-Store

Table 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 the Lewis and Reiley (2014) experiment with the same retailer that found that in-store sales represented 85% of the total treatment effect.

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

A.2.3 Probability of Purchase Versus Basket Size

Marketers often decompose the effect of ads on sales into increasing the probability of purchase and buying more per shopping trip. We examine the experimental differences in probability of purchase, the number of shopping trips, and the 'basket size' or purchase amount conditional on a transaction. We present the basket size results as descriptive since we cannot separately identify the basket size of marginal consumers (those for whom ad exposure caused them to make one or more purchases instead of zero) from those who would have made at least one purchase anyway. To examine the impact on shopping trips, we construct a variable equal to the number of days in which a consumer made a transaction during the campaign period.[13] 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 7 illustrates our results. The first column restates our original results for total sales. The second column presents results of a linear-probability regression for a transaction indicator dummy variable. The probability of a transaction increases with advertising by 0.43% (s.e. 0.46%) for the Full treatment group and by 0.47% (s.e. 0.46%) for the Half treatment group, relative to a baseline purchase amount of 7.7% for all treated consumers in the sample, though the increases are not statistically significant.[14]

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. Relative to a baseline mean basket size of $171, these represent percentage increases of 2.24% and 0.74% respectively.

Table 7's column (4) shows the impact of ads on shopping trips. The Full treatment produces 0.0020 additional trips (p=0.013) and the Half treatment produces 0.0011 additional trips (p=0.14) per person. These point estimates represent 1,092 incremental transactions in the Full group and 640 in the Half group. The additional columns of the table show that the effects are larger for in-store ($p < 0.1$) than for online sales ($p < 0.01$). Because the mean number of channel-day transactions per person is 0.112, the Full treatment effect represents a 1.8% increase in total transactions. This represents half of the 3.6% total treatment effect on sales. In contrast, Lewis and Reiley (2014) found that increased probability of purchase represents only around one-quarter of the total effect on purchases. However, their data only allows them to examine the impact on the probability of any transaction during the campaign and misses the potential role of multiple

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

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

Table 5: Effects of the Advertising During and After the Campaign

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

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

shopping trips in the ad effect.

A.3 Proximity Results Robustness

In our main results, we present evidence that shows that the consumer proximity to the store predicts ad effectiveness. Towards this, we consider five measures that differentiate consumers: household distance to the nearest store, transaction recency, frequency, monetary value and household income. In this online appendix, we examine the robustness of these results to different levels of proximity. For instance, our results contrast the effect of sales among consumers who live within and beyond a 1 mile radius; in the appendix, we consider multiple radii of different length. We examine the effect on channel-day shopping trips by consumer proximity (available upon request), which yields similar results. Given the low-power setting—particularly in the Half treatment group—the results are quite robust.

Table 6: Effects of the Advertising, Online versus Offline

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

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

Distance to Nearest Store

Table A.1 illustrates the effect of sales among consumers who live within and outside various radii of different length: 0.5, 1, 1.5, 2.5, 3, 4, and 5 miles. The results for 1 mile are robust for the radius 2 miles in that an F-test of inequality has $p = 0.099$ for the Full group and the joint inequality F-tests has $p = 0.108$ for the Full and Half groups. The 1 mile and 2 mile radii encompass 3.4% and 10.8% respectively of the sample for which we have distance data. For radii larger than 3 miles, the point estimates in the Full and Half groups are larger outside the radii; nevertheless, these differences are neither individually nor jointly significant.

Recency

Our main results consider whether a consumer transacts within the 8 weeks prior to the beginning of the experiment plus the time between the experiment's beginning and the day of the user's first ad exposure in the experiment. Table A.2 demonstrates robustness as we vary the definition of

Table 7: Effects of the Advertising: Probability of Purchase, Basket Size, versus Shopping Trips

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

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

recency to include 2 weeks, 4 weeks, 6 weeks, 8 weeks, and 44 weeks prior. We construct this using the retailer sales data which provides weekly sales for weeks 1-8 prior to the experiment and a single variable for the weeks 9-44 sales total. The results are very robust. For the Full group, the recent transactors have higher sales lift, and a F-test of inequality is significant at 10% for 4 weeks and up and at 5% for 6 weeks and up. The lift among recent transactors in the Half group is similarly significant at the 5% level of those when defined as transacting within 44 weeks. The joint test of inequality for the Full and Half groups is significant at 10% for 4 weeks and up and at 5% for weeks 6 and up. As we widen the definition of recency, the ‘near’ sample size increases, which boosts the power of the test to detect inequalities. Furthermore, the recently transacted consumers always have higher point estimates than those who have not transacted recently. The point estimates of the recently transacted consumers in the Full and Half groups broadly fall as we widen the definition of recency, with only a couple of exceptions.

Frequency

We compare the ad lift by user purchase frequency by whether the user belongs to the top three of five retailer frequency categories. Table A.3 shows that this is robust to splitting ad lift according to the top four frequency categories (72% of users) as both the Full group and joint tests of inequality are significant at 10%. The top two groups only contain 11% of users and top group is only 5%. Though the experimental lift estimates are higher within the top and top two frequency categories, the inequality tests are not significant at 10%.

Monetary Value

Whereas our main results present the effect of ads on sales by whether a user has spent $1,000 at the retailer in the previous two years, Table A.4 considers alternate cutoffs. Specifically, Table A.4 compares users by their monetary value as defined by whether their sales at the retailer in the previous years exceeds the following cut-offs: $2,000, $1,500, $1000, $750, $500, $250. An F-test of inequality between the sales lift high and low monetary value users rejects that the low value consumers have higher lift at the 5% level for the cut-offs between \$1000 and \$250 in the Full group. The joint F-test of inequality is significant at 10% for these cut-offs and significant at 5% for the cut-offs between \$750 and \$250. The point estimates for the high monetary value users are higher in all cases but for the imprecisely estimated Half treatment users who spend over \$2,000.

Income

The paper splits sales lift by whether the user's annual household income is more or less than \$100,000. In Table A.5, we consider alternate cutoffs of \$250,000, \$150,000, \$100,000, \$75,000, and \$50,000. At the outset, we should point out that the relationship between income and taste proximity to the retailer is likely non-monotonic: the store does not cater to the super-rich or to lower income users. In the Full group, the point estimates of the sales lift among the wealthier users are all higher and they decrease monotonically as we relax our wealth cut-off. For a cutoff of \$250,000, the F-test of inequality is significant at 10% in the Full group. For cut-offs between \$75,000 and \$150,000, the test of inequality is significant at 5%. In the Half treatment group, the picture is murkier since the point estimate in the wealthy group is lower for all but the \$100,000 cut-off group though the F-test of equality here cannot be rejected at the 15% level in all cases.

Table A.1: Heterogenous Effects on Sales by Consumer's Crow Flies Distance to Retailer's Nearest Store

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Condition	Lives within 0.5 mile of nearest store	Lives within 1 mile of nearest store	Lives within 1.5 mile of nearest store	Lives within 2 mile of nearest store	Lives within 2.5 mile of nearest store	Lives within 3 mile of nearest store	Lives within 4 mile of nearest store	Lives within 5 mile of nearest store
Subset of Users[a]	Treated	Treated	Treated	Treated	Treated	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x	x	x	x	x	x
`Near' in Full Treatment	6.012	2.883**	1.857*	1.535*	0.720	0.397	0.280	0.445
	(3.840)	(1.430)	(1.082)	(0.805)	(0.652)	(0.558)	(0.448)	(0.400)
`Far' in Full Treatment	0.517**	0.485**	0.472**	0.449*	0.538**	0.609**	0.686**	0.641**
	(0.230)	(0.233)	(0.235)	(0.239)	(0.245)	(0.252)	(0.268)	(0.280)
`Near' in Half Treatment	-0.596	1.523	0.578	0.816	0.521	-0.118	0.343	0.476
	(3.765)	(1.528)	(1.142)	(0.832)	(0.676)	(0.573)	(0.457)	(0.404)
`Far' in Half Treatment	0.389	0.340	0.366	0.327	0.354	0.505*	0.396	0.322
	(0.237)	(0.240)	(0.241)	(0.246)	(0.252)	(0.260)	(0.276)	(0.292)
Covariates: Full Set[c]	x	x	x	x	x	x	x	x
H_0 : $\Delta sales_{near} < \Delta sales_{far}$								
Full: F-test (p-value)	0.077	0.049	0.106	0.099	0.398	-	-	-
Half: F-test (p-value)	-	0.223	0.428	0.287	0.409	-	-	-
Full & Half: F-test (p-value)	-	0.064	0.096	0.108	0.240	-	-	-
H_0 : $\Delta sales_{near} = \Delta sales_{far}$								
Full: F-test (p-value)	0.153	0.0979	0.212	0.197	0.795	0.729	0.437	0.688
Half: F-test (p-value)	0.794	0.445	0.856	0.573	0.818	0.322	0.921	0.758
Full & Half: F-test (p-value)	0.150	0.254	0.384	0.431	0.960	0.604	0.701	0.781
% Non-missing condition data[d]	77.1%	77.1%	77.1%	77.1%	77.1%	77.1%	77.1%	77.1%
`Near' proportion	0.9%	3.4%	6.8%	10.8%	15.4%	20.1%	29.6%	38.0%
Observations	1714704	1714704	1714704	1714704	1714704	1714704	1714704	1714704
R-squared	0.103	0.103	0.103	0.103	0.103	0.103	0.103	0.103

Average effect of Treatment on the Treated estimates conditional on the relevant consumer characteristics. Dependent variable is sales during the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details). We also include indicator variables for the given condition. [d]The data for some variables (distance, income, gender) is incomplete. We include all observations however as these improve the estimates for the covariates.

Table A.2: Heterogenous Effects on Sales by Consumer Purchase Recency

	(1)	(2)	(3)	(4)	(5)
Condition	Transacted in 2 weeks pre-treatment	Transacted in 4 weeks pre-treatment	Transacted in 6 weeks pre-treatment	Transacted in 8 weeks pre-treatment	Transacted in 44 weeks pre-treatment
Subset of Users[a]	Treated	Treated	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x	x	x
`Near' in Full Treatment	1.485	1.918*	1.922**	1.624**	0.721**
	(1.434)	(1.046)	(0.876)	(0.752)	(0.307)
`Far' in Full Treatment	0.366**	0.215	0.130	0.129	0.0168
	(0.158)	(0.149)	(0.140)	(0.139)	(0.121)
`Near' in Half Treatment	1.283	1.085	1.140	0.804	0.418
	(1.511)	(1.091)	(0.903)	(0.776)	(0.314)
`Far' in Half Treatment	0.105	0.0681	0.00430	0.0485	-0.150
	(0.155)	(0.147)	(0.140)	(0.138)	(0.111)
Covariates: Full Set[c]	x	x	x	x	x
H_0 : $\Delta sales_{near} < \Delta sales_{far}$					
Full: F-test (p-value)	0.219	0.054	0.022	0.025	0.017
Half: F-test (p-value)	0.219	0.178	0.107	0.169	0.044
Full & Half: F-test (p-value)	0.169	0.068	0.032	0.037	0.020
H_0 : $\Delta sales_{near} = \Delta sales_{far}$					
Full: F-test (p-value)	0.438	0.107	0.0435	0.0509	0.0331
Half: F-test (p-value)	0.438	0.356	0.214	0.338	0.0875
Full & Half: F-test (p-value)	0.675	0.272	0.127	0.148	0.0801
% Non-missing condition data[d]	100%	100%	100%	100%	100%
`Near' proportion	10.3%	15.4%	19.5%	23.2%	65.3%
Observations	1714704	1714704	1714704	1714704	1714704
R-squared	0.104	0.103	0.103	0.103	0.103

Average effect of Treatment on the Treated estimates conditional on the relevant consumer characteristics. Dependent variable is sales during the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details). We also include indicator variables for the given condition. [d]The data for some variables (distance, income, gender) is incomplete. We include all observations however as these improve the estimates for the covariates.

Table A.3: Heterogenous Effects on Sales by Consumer Frequency

	(1)	(2)	(3)	(4)
Condition	Frequency Category 1	Frequency Category 1-2	Frequency Category 1-3	Frequency Category 1-4
Subset of Users[a]	Treated	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x	x
`Near' in Full Treatment	2.185	1.889	1.189**	0.606**
	(3.343)	(1.565)	(0.567)	(0.277)
`Far' in Full Treatment	0.396***	0.305**	0.111	0.147
	(0.144)	(0.128)	(0.107)	(0.161)
`Near' in Half Treatment	2.152	0.537	0.425	0.311
	(3.488)	(1.628)	(0.580)	(0.284)
`Far' in Half Treatment	0.129	0.183	0.117	-0.0101
	(0.140)	(0.123)	(0.104)	(0.145)
Covariates: Full Set[c]	x	x	x	x
H_0 : $\Delta sales_{near} < \Delta sales_{far}$				
Full: F-test (p-value)	0.297	0.157	0.031	0.076
Half: F-test (p-value)	0.281	0.414	0.301	0.155
Full & Half: F-test (p-value)	0.204	0.141	0.039	0.084
H_0 : $\Delta sales_{near} = \Delta sales_{far}$				
Full: F-test (p-value)	0.593	0.314	0.0620	0.151
Half: F-test (p-value)	0.562	0.828	0.601	0.309
Full & Half: F-test (p-value)	0.815	0.564	0.154	0.336
% Non-missing condition data[d]	100%	100%	100%	100%
`Near' proportion	4.6%	10.9%	33.9%	72.0%
Observations	1714704	1714704	1714704	1714704
R-squared	0.103	0.103	0.103	0.103

Average effect of Treatment on the Treated estimates conditional on the relevant consumer characteristics. Dependent variable is sales during the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details). We also include indicator variables for the given condition. [d]The data for some variables (distance, income, gender) is incomplete. We include all observations however as these improve the estimates for the covariates. The frequency variables are derived from the

Table A.4: Heterogenous Effects on Sales by Consumer's Monetary Value

	(1)	(2)	(3)	(4)	(5)	(6)
Condition	Spent over $2,000 in 2 years pre-treatment	Spent over $1,500 in 2 years pre-treatment	Spent over $1,000 in 2 years pre-treatment	Spent over $750 in 2 years pre-treatment	Spent over $500 in 2 years pre-treatment	Spent over $250 in 2 years pre-treatment
Subset of Users[a]	Treated	Treated	Treated	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x	x	x	x
`Near' in Full Treatment	6.012	2.883**	1.857*	1.535*	0.720	0.397
	(3.840)	(1.430)	(1.082)	(0.805)	(0.652)	(0.558)
`Far' in Full Treatment	0.517**	0.485**	0.472**	0.449*	0.538**	0.609**
	(0.230)	(0.233)	(0.235)	(0.239)	(0.245)	(0.252)
`Near' in Half Treatment	-0.596	1.523	0.578	0.816	0.521	-0.118
	(3.765)	(1.528)	(1.142)	(0.832)	(0.676)	(0.573)
`Far' in Half Treatment	0.389	0.340	0.366	0.327	0.354	0.505*
	(0.237)	(0.240)	(0.241)	(0.246)	(0.252)	(0.260)
Covariates: Full Set[c]	x	x	x	x	x	x
$H_0 : \Delta sales_{near} < \Delta sales_{far}$						
Full: F-test (p-value)	0.129	0.175	0.047	0.050	0.046	0.050
Half: F-test (p-value)	-	0.483	0.368	0.434	0.479	0.450
Full & Half: F-test (p-value)	-	0.141	0.051	0.047	0.039	0.046
$H_0 : \Delta sales_{near} = \Delta sales_{far}$						
Full: F-test (p-value)	0.257	0.349	0.094	0.099	0.092	0.100
Half: F-test (p-value)	0.899	0.965	0.736	0.868	0.957	0.899
Full & Half: F-test (p-value)	0.366	0.562	0.204	0.187	0.155	0.183
% Non-missing condition data[d]	100%	100%	100%	100%	100%	100%
`Near' proportion	10.0%	13.7%	20.1%	25.5%	34.2%	51.5%
Observations	1714704	1714704	1714704	1714704	1714704	1714704
R-squared	0.103	0.103	0.103	0.103	0.103	0.103

Average effect of Treatment on the Treated estimates conditional on the relevant consumer characteristics. Dependent variable is sales during the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details). We also include indicator variables for the given condition. [d]The data for some variables (distance, income, gender) is incomplete. We include all observations however as these improve the estimates for the covariates.

Table A.5: Heterogenous Effects on Sales by Consumer Income

	(1)	(2)	(3)	(4)	(5)
	Earns	Earns	Earns	Earns	Earns
	$250,000	$150,000	$100,000	$75,000 or	$50,000 or
Condition	or more	or more	or more	more	more
Subset of Users[a]	Treated	Treated	Treated	Treated	Treated
Sales After First Ad Exposure[b]	x	x	x	x	x
\`Near' in Full Treatment	1.785*	1.032**	0.810**	0.706***	0.490**
	(1.017)	(0.457)	(0.332)	(0.267)	(0.227)
\`Far' in Full Treatment	0.332*	0.233	0.112	-0.0603	0.345
	(0.200)	(0.219)	(0.240)	(0.295)	(0.452)
\`Near' in Half Treatment	-0.953	0.141	0.383	0.151	0.134
	(0.949)	(0.472)	(0.339)	(0.271)	(0.232)
\`Far' in Half Treatment	0.346*	0.254	0.0463	0.381	0.836*
	(0.209)	(0.221)	(0.242)	(0.305)	(0.460)
Covariates: Full Set[c]	x	x	x	x	x
H_0 : $\Delta sales_{near} < \Delta sales_{far}$					
Full: F-test (p-value)	0.081	0.057	0.045	0.028	0.388
Half: F-test (p-value)	-	-	0.210	-	-
Full & Half: F-test (p-value)	-	-	0.059	-	-
H_0 : $\Delta sales_{near} = \Delta sales_{far}$					
Full: F-test (p-value)	0.161	0.114	0.090	0.055	0.775
Half: F-test (p-value)	0.181	0.828	0.420	0.572	0.173
Full & Half: F-test (p-value)	0.023	0.163	0.237	0.031	0.226
% Non-missing condition data[d]	98.2%	98.2%	98.2%	98.2%	98.2%
\`Near' proportion	9.5%	31.2%	52.6%	70.1%	87.8%
Observations	1714704	1714704	1714704	1714704	1714704
R-squared	0.103	0.103	0.103	0.103	0.103

Average effect of Treatment on the Treated estimates conditional on the relevant consumer characteristics. Dependent variable is sales during the two weeks of the experiment. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. [a]Treated users are those who are exposed to either the retailer or the control ad. [b]Sales after the first ad exposure modifies the outcome measure to exclude all sales prior to a user's first exposure to either the retailer or control ad. [c]Includes demographics, customer categories, two-year of past sales, and exposure intensity (see Table 2 for details). We also include indicator variables for the given condition. [d]The data for some variables (distance, income, gender) is incomplete. We include all observations however as these improve the estimates for the covariates.