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

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

June 2, 2016

Abstract

We examine a large-scale field experiment that measures the effect of online display advertising. Yahoo! Research and an American apparel retailer collaborated on the experiment that tracked the online and offline sales of 3 million retailer customers. We measure the impact of the 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 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>, Netflix. Reiley: <david@davidreiley.com>, Pandora Media, Inc. and University of California at Berkeley. The authors performed this research while employed at Yahoo! Research. We acknowledge Yahoo! both for its financial support and for its support of unbiased scientific research through its ex-ante decision to allow us to publish the results of the experiment no matter how they turned out. We would like to thank Valter Sciarillo, Jennifer Grosser Carroll, Taylor Schreiner, Eddie Babcock, Iwan Sakran, and many others at Yahoo! and at our client retailer who made this research possible. We thank Eric Anderson, Susan Athey, Mitch Lovett, Ben Jones, Preston McAfee, Rob Porter, Justin Rao, Elie Tamer, and Florian Zettelmeyer for helpful discussions and Mary Kraus for proofreading.

1 Introduction

Advertisers assume that diminishing returns set in once a consumer has seen enough ads, but advertisers disagree over what constitutes 'enough.' Advertisers also disagree over whether advertising should target a firm's best customers, or whether it should instead attempt to convert marginal customers into more lucrative customers. In this paper, we shed light on these issues by revisiting the large-scale controlled experiment of online display advertising on Yahoo! presented by Johnson et al. (2016a). This experiment examines two consecutive weeklong ad campaigns that target three million of the retailer's existing customers. The experiment stands out for its large scale and in its combination of consumer-level ad exposure with retail purchase data. Whereas Johnson et al. (2016a) compare approaches for improving measurement efficiency in an experiment, we go further by examining heterogeneous treatment effects of interest to marketers. On the question of ad frequency, we experimentally vary the intensity of treatment and find that 'enough' is higher than many expect: the returns to advertising appear constant even out to 50 ad exposures in two weeks. We also show that the retailer's ads generate the largest sales lift among the retailer's best customers and those customers who live near a store. Thus, as the pun in our title suggests, both repetition and proximity (in geographic or abstract product space) can produce benefits in advertising.

Our experiment combines an ad weight test with a holdout test in order to examine the shape of the consumer ad frequency response curve. The experiment randomizes users into three treatment groups of equal size. The 'Half' and 'Full' treatment groups reveal the effect of doubling the average ad weight from 16.7 retailer impressions to 33.4 impressions per consumer. The 'Control' group is held out from seeing the retailer ads and serves to establish the baseline sales had the retailer not advertised. Field experiments are the gold standard for measuring the causal effect of online advertising (Lavrakas, 2010). This view is supported by Gordon et al. (2016) who show that observational ad measurement approaches used in industry are unable to reproduce the causal effects of advertising measured in field experiments with twelve different online display ad campaigns.

One challenge of measuring ad effectiveness is low statistical power. Lewis and Rao (2015)

discuss this challenge: the effects of economically profitable online display ad campaigns ads are plausibly orders of magnitude smaller than the variation in the marketing outcome data. Consequently, unbiased field experiments can require millions of consumer-week observations to yield statistically significant results. For this reason, Johnson et al. (2016a) doubled the sample size of an earlier experiment with the same retailer examined by Lewis and Reiley (2014), improving the experimental design and collecting data designed to increase statistical power further. This powerful experiment enabled Johnson et al. (2016a) to detect statistically significant effects of advertising, and will allow us in the present paper to tackle new research questions about heterogeneous treatment effects like the effects of increased frequency.

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 approximately 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. To further establish that the campaign is profitable, we impose and test restrictions on the functional form of the effect of ad frequency.

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 (2015) experimentally varies ad frequency on a restaurant search website and finds decreasing returns to banner-style search ads after three exposures in a session. The results on returns to scale might depend on differences in purchase intent: Sahni (2015) 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 taste proximity, in abstract product-preference space, to 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), which 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. The third section provides descriptive statistics. The fourth section presents our measurements of the causal effects of the advertising, focusing on the effects of frequency and heterogeneous treatment effects by consumer proximity. The final section concludes.

2 Experimental Design

Johnson et al. (2016a) discuss the experimental design in detail. Below, we review some critical elements.

The retailer's online display advertising features two consecutive week-long campaigns. Both campaigns are image advertising, featuring the retailer's brand, the apparel manufacturer's brand, and a product from this manufacturer. The creatives are so-called 'flash' ads that feature slideshow transitions between images. Our retailer prefers to remain anonymous, but Figure 1 shows a similar

creative from retailer Target and manufacturer Missoni—neither of which participated in this study. By contrast with this example ad, the creatives in our experiment feature the retailer's brand at least as prominently as the manufacturer's brand. As in Figure 1, the retailer ads do not promote a sale nor do they include price information. The ads appear throughout Yahoo! properties including Mail, News, and Autos. The week-one campaign rotates three creatives: a woman modelling a jacket, a man modelling a polo shirt, and a woman's shoe. The week-two campaign features another woman modeling a dress. We do not re-randomize users between campaigns nor do we assign users to the different creatives in week two. Consequently, the experimental design cannot speak to the effect of creative rotation on the ad frequency curve.

The study employs control ads with an orthogonal call to action, promoting Yahoo!'s search engine. The Control group receives the same total number of control ad impressions as the Full group receives in retailer impressions. An intermediate Half group also receives the same number of total impressions in our experiment, but these are divided in half, with equal numbers of impressions for the retailer and for Yahoo! Search. This design means that, for each experimental impression, a Half group user sees either a retailer ad or a control impression with equal probability. As Johnson et al. (2016a) show, the control ads together with the retailer ads serve to identify which users are exposed to at least one experimental impression, as well as the first day of exposure for each user. Note that only 55% of the study's three million subjects are exposed to an experimental ad, because some subjects do not browse sufficient pages on Yahoo! properties to see one of the ads in our experiment. The sum of the retailer and control impressions seen by a given consumer serve to identify the number of retailer ads that person would have seen if they had been in the Full treatment group. For the control ads to be valid, they must be delivered the same as the focal advertiser's ad. Johnson et al. (2016b) explain that this may not hold in general, especially for click- or conversion-optimized campaigns. We set up the control ads to avoid these concerns, avoiding the optimization of the campaign's clicks or conversions. In the next section, we show that the impressions are distributed equally across treatment groups.

In the experiment, the retailer targets its existing customers using a 'database-match' campaign.

In a database-match campaign, the advertiser and the ad platform use a third party to match consumers in each firm's customer database. The match returned about 3 million users who have both purchased at the retailer in the past and signed up with Yahoo!. This match reveals the user-level sales not only at the retailer's online store but also the offline sales at their brick-and-mortar locations. Such sales data remain rare in the online advertising literature and are important here because the majority of the apparel retailer's sales are in-store. In addition, the retailer shared two years of user-level sales history, which allows us to look for heterogeneous treatment effects by a user's purchase Recency, Frequency and Monetary Value (RFM, see Section 4.3).

One important advantage of our design is that we assign treatment to a logged-in user rather than to a browser cookie. This means that a logged-in user will be assigned to the same treatment group across the user's different browsers and devices, each of which has its own cookie. Cookie-level randomization attenuates estimates of the causal effect of advertising. To see this, suppose each person has two cookies and each cookie is assigned with equal probability to a treatment or control group. Then a quarter of people will be assigned to the treatment group uniquely, a quarter to the control group uniquely, but half of people will have a cookie in each group. Thus, the cookie-level experimental difference only arises in expectation from people in a single treatment group and the average difference will be diluted by people in both groups. Coey and Bailey (2016) discuss this issue in detail and use a real-world example to suggest that cookie-level randomization could understate the person-level lift by as much as a factor of three. Nevertheless, person-level studies remain the exception in field experiments (see e.g. Gordon et al. 2016) as most platforms are restricted to cookie-level studies (see e.g. Lambrecht and Tucker 2013; Hoban and Bucklin 2015; Johnson et al. 2016b). Smaller issues, like device sharing and failure to correctly attribute each purchase to the correct customer, will also attenuate the lift estimates, as Johnson et al. (2016a) explain. Still, our experiment has significant advantages over those that fail to measure offline purchases or that assign treatment based on cookies.

3 Data

In this section, we summarize the data, supplement the randomization checks presented in Johnson et al. (2016a), and discuss the location data.

Table 1 summarizes user-level variables for the 55.4% of users who are exposed to at least one experimental impression. The exposed users are exposed to an average of 33.4 experimental ads during the two weeks of the campaign, with the Half group seeing 16.7 retailer ads. In the Full group, the click-through rate on the retailer ads was 0.19% with 4.9% of exposed users clicking on a retailer ad. In the two weeks prior to the experiment, average sales at the retailer were $19.12 and were highly variable with a standard deviation of $149. 70% of exposed users are women and the average age is 44 years. On average, exposed consumers live 21 miles from the nearest retailer location, but half live within 6.6 miles.

Valid experiments require that the experimental treatment assignment is independent of the subject's potential attributes. Johnson et al. (2016a) verify this by testing and failing to reject the equality of average user characteristics, including pre-treatment sales outcomes. Johnson et al. (2016b) explain that ad platforms often deliver the retailer and control ads differently and expose different user types, which would invalidate the experiment. To that end, in Figure 2 we add a randomization check for the distribution of experimental ads across treatment groups, which also reveals no significant differences.

In Table 1, we present a summary of additional comparisons of means between the two groups that were not previously presented in Johnson et al. (2016b). The first section of the table compares user attributes for the subset of users who were exposed to the campaign (statistics for the full set of eligible users were presented in Johnson et al. (2016b)). The second section of the table presents comparisons of pre-campaign purchase amounts, two years and two weeks before the campaign, respectively. These tests collectively fail to reject the hypothesis of equality of means for user characteristics, implying that no selection bias was accidentally introduced by the ad-serving algorithm.

A key methodological innovation of Johnson et al. (2016a) is to increase measurement precision

by trimming outcome data that never could have been caused by the advertising. In particular, the campaign cannot affect a user's outcome variable until after the user is exposed to the first ad impression, so pre-exposure outcomes only contribute noise to the estimator. We therefore improve statistical precision by trimming from our dataset any purchases made during the campaign but before a consumer's first ad exposure. For this approach to be valid without biasing the estimates, a Full group user who is not exposed until the dth day of the campaign must have a counterpart in the control group who is not exposed until the dth day. We therefore provide, in the second and third sections of Table 1, two more randomization checks that test for differences by exposed users across different treatment groups: the pre-exposure, in-campaign average sales ($2.31) and the average day of the first experimental ad exposure (3.28 days). Though we see no significant differences for pre-exposure, in-campaign average sales (p-value=0.37), we do see significant differences for the day of the first exposure ($p = 0.006$).

The differences in the day of first exposure may result from the ad platform's pacing algorithm but do not meaningfully impact our results. Presenting multiple validation checks creates a multiple hypothesis testing problem: some tests will reject due to random chance. Nonetheless, we hypothesize that these differences could result from the ad platform's budget pacing algorithm, which operates independently on each treatment group. Budget pacing algorithms smooth out the delivery of an ad campaign over time, so that even identical campaigns can differ slightly with the arrival of users to the platform.[1] Though the differences in the first exposure date are statistically different, the absolute difference is small (< 0.02 days, or less than half an hour out of a two-week campaign period). The validation checks for the number of impressions and the pre-exposure, in-campaign sales between treatment groups reassure us that the first day of exposure differences do not invalidate our lift estimates that trim sales prior to the first exposure day.[2]

In Section 4.3, we consider heterogeneous treatment effects by the user's physical proximity to the nearest brick-and-mortar store. To this end, we add geographic location data to the data of

[1] This issue can be resolved by Ghost Ads and related methodologies proposed by Johnson et al. (2016b).

[2] We test for difference in our trimmed outcome lift estimates with a Hausman test. The comparison between the lift estimates in column (2) and column (3) of Table 2 reveal no significant differences ($p = 0.44$).

Johnson et al. (2016a). To compute proximity, we use nine-digit zip code data from Yahoo! and a third-party data broker, which are available for 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 3 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.[3]

4 Results

The results section is divided into three subsections, which we briefly preview. Section 4.1 reviews the experimental estimates from Johnson et al. (2016a) for the sales lift during the two weeks of the ad campaign. Our preferred experimental estimates indicate 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. 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.

4.1 Overall Campaign Impact

We begin by reviewing the ad effect estimates during the two weeks of the campaign. Johnson et al. (2016a) present these estimates and describe how to use the control ads and covariate data to improve the precision of the estimates.

[3]The haversine formula calculates the distance between two pairs of latitude and longitude coordinates while correcting for the spherical curvature of the planet. Note that the results are similar if we use driving distance instead.

In Table 2, we present Average Treatment Effect on the Treated (ATET) estimates for the effect of the ads on sales. Table 2 is an abbreviated version of Table 2 from Johnson et al. (2016a) and builds to our preferred estimator. Column (1) presents the least precise estimates (standard error of \$0.31), which do not use the control ads. Column (1) rescales the Intent to Treat estimates by the inverse exposure probability (55.4%) to recover ATET (see ch. 7 of Manski 2007 for the derivation). Column (2) uses the control ads to condition on exposed users across all treatment groups; trimming unexposed users shrinks the standard errors by 25%. An innovation of Johnson et al. (2016a) was to trim the pre-exposure, in-campaign sales from the outcome variable. We explain the logic behind this innovation in Section 3. Column (3) shows that this innovation reduces standard errors by a further 8%. Finally, our preferred estimator in column (4) includes 236 covariate variables that explain some of the variation in the sales outcome variable ($R^2 = 0.09$), which reduces the standard errors by another 5%. These user-level covariates include demographics, two-years of granular sales history, retailer-defined RFM customer categories, and ad-exposure intensity indicators. In all, these approaches improve the precision of the estimates by 34%. This is equivalent to increasing the sample size and advertising budget by 80%. These statistical power gains will allow us to examine heterogeneous treatment effects that cut the sample into subgroups.

Our preferred estimates indicate that sales are 3.6% higher in the Full group and 1.7% higher in the Half group than the baseline established by the Control group. The lift in the Full group is \$0.477 (standard error \$0.204), which rejects the null of a zero sales effect ($p = 0.020$, two-sided). The Half group lift is \$0.221 (s.e. \$0.209), however this is not statistically significant ($p = 0.289$, two-sided). The two estimates are marginally significant in a joint F-test ($p = 0.065$, two-sided). To give the lift estimates context, the average advertising cost is \$0.17 in the Full group.

To calculate the campaign's return on investment, we assume that the retailer's gross margin is 50% following Johnson et al. (2016a). The return on investment for the retailer's \$132,000 total spend is then 51% with 95% confidence intervals of [-101%, 204%]. This level of uncertainty is disconcerting giving the sample size and ad spend involved, but reflects the managerial challenge of this low-powered environment, as documented by (Lewis and Rao, 2015). In Section 4.2, we

show that we can make and test assumptions on the ad frequency curve that yield statistically significant evidence that the campaign generated a positive return on investment and thereby improve managerial decision-making.

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 4. First, the three treatment groups provide exogenous variation in frequency. Second, the Half treatment group provides additional exogenous variation in frequency within each ad type. 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 illustrated in Figure 4. 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.

We first 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 exposed subjects. As Figure 2 illustrates, data on users who see more than 50 impressions 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 provide estimates for $\theta > 50$ in Section 4.2.3 below.

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 nonlinearities in Section 4.2.2, we will be unable to reject 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 imply that ad type plays a secondary role for $0 < \theta \leq 50$. We say this because the coefficient on ad type is small and insignificant. We cannot reject the null hypothesis that the curve $\gamma(\theta) = \gamma_0 + \gamma_1\theta$ has slope equal

to zero.[4] Figure 5 nonparametrically estimates the relationship between ad type θ and indexed sales among exposed users. Figure 5 shows that sales both during the experiment (for users with $f = 0$ in the Control and Half groups) and 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 all the variation in 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 this assumption 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 decrease discontinuously from $\theta = 0$ to $\theta = 1$ among unexposed users.[5] In other words, a non-experimental comparison of exposed versus unexposed users would produce substantial downward bias in the estimated causal effects of advertising. 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 (see Section 4.2.3).

Assuming $\gamma_1 = 0$ improves our precision with minimal introduction of 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

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

[5] Relative to unexposed users, users in the Half and Control groups with $f = 0$ and $\theta = 1$ have \$0.26 lower sales and those with $f = 0$ and $0 < \theta \leq 50$ have \$0.29 lower sales during the campaign.

generates revenues 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 assumed gross margin of 50%. 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 the hypothesis that the ads reduced sales for that group. 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 statistically 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 (2015) 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. Experiments provide exogenous variation in advertising exposure, allowing us both to improve precision directly (with variation that would not have occurred naturally) and to check modeling assumptions that, if true, can improve precision further. For example, in this setting our control group shows us we can safely assume away the correlation between ad type and purchase amount for $0 < \theta \leq 50$, but we cannot safely assume that people with zero potential exposures ($\theta = 0$) purchase equivalent amounts to those with positive potential exposures ($0 < \theta \leq 50$). The experimental and modeling approaches complement each other in the challenging task of optimizing ad spend.

4.2.2 Constant Returns to Frequency

Earlier, we assumed constant returns to frequency in order to get more precise estimates of the marginal effects of advertising. How reasonable is this linearity assumption? Figure 6 shows

nonparametric estimates of the effects of frequency, using the model $Y = \beta_f \cdot f + \gamma_\theta + \epsilon$. 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 $f = 0$. Figure 6 also includes a line with the slope of our linear marginal frequency estimate $3.7¢$ that intersects with almost all the confidence intervals of the $\beta_{f'}$ estimates. Note that this model exploits both exogenous and endogenous variation in frequency, making the assumption (validated above and again below) that for consumers with $0 < \theta \leq 50$, control-group baseline sales are approximately constant in θ. Though the nonparametric fit is noisy, we see that our linearity assumption above is quite reasonable.

Next, we impose a quadratic frequency model and ask whether we can detect any deviation from constant returns to frequency in the range $0 < \theta \leq 50$. We do so by estimating the 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. 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 with constant returns to frequency is a reasonable assumption for us to have made in our earlier results.

In our experiment, we find approximately constant returns to advertising out to 50 impressions over the course of two weeks, a rather high figure. This result is difficult to compare to the existing experimental literature on ad frequency. Lewis (2010)'s meta-analysis estimates frequency curves for up to 50 impressions on a single day of advertising on Yahoo!'s front page using a natural experiment giving binomial variation in frequency alone. Using clicks as the outcome, he finds that a minority of campaigns—a third—demonstrate constant returns for clicks up to 20 exposures per day and finds some campaigns with constant returns up to 50 impressions. Nevertheless, the majority of the campaigns with decreasing returns showed only 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 (2015) 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 (2015), the ad exposures in our experiment on Yahoo! are unrelated to purchase intent, and thus more likely to be ignored as irrelevant. 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 on consecutive weeks, and the first campaign includes three kinds of ad creatives.

4.2.3 Frequency Model Robustness

In this section, we perform two more robustness checks. We consider a more general model that does not pool together users with different ad types and present lift estimates for users with $\theta > 50$.

First, we ask whether the effect of frequency is heterogeneous for consumers with different levels of browsing behavior in the range $0 < \theta \leq 50$. We consider a generalized model of the form $Y = \beta_\theta \cdot f + \gamma_\theta + \epsilon$, allowing the returns to frequency f to vary nonparametrically 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 consumer type 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. Because we estimate a separate model for each θ, here we rely exclusively on exogenous, experimental variation in frequency. Figure 7 collects the 50 slope parameter estimates β_θ for $0 < \theta \leq 50$. We see that these parameter estimates 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. This tells us that our earlier

model pooling consumers with different θ is not at all unreasonable. We find it instructive to note that although lower θ have more observations, the confidence intervals are tighter for higher θ because they contain greater variation in f, providing greater leverage for estimating its marginal effects. The estimates exhibit no systematic tendency for consumers with either higher or lower amounts of browsing θ) to have significantly higher returns to advertising than other consumers do. Also, 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 statistical power problem for this and other models without parametric assumptions imposed.

Finally, 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 arbitrary, but 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 exposed 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 exposed 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 purchase 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 exposed 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, 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.[6] 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

[6] 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 is said to have remarked: "Half the money I spend on advertising is wasted; the trouble is I don't know which half." In this paper, we demonstrate progress in finding the "wasted half" through measuring heterogeneous treatment effects in a controlled experiment. The experiment enables us to tackle a classic endogeneity problem in marketing, the effect of advertising on demand, by introducing our own exogenous variation in advertising exposure.

Due to issues of statistical power in measuring a single advertising campaign (see Lewis and Rao 2015), Wanamaker's lament remains more of a problem than some might expect. Even with a million subjects per treatment group, our modern-day retailer's experiment does not detect a statistically significant ad effect at normal exposure intensity (i.e., the Half treatment) even though the sales lift was profitable. However, our results on frequency should give Mr. Wanamaker hope: because the effects appear to be approximately linear for at least the first 50 online display impressions per person, we can improve statistical power by increasing the frequency of a campaign. We find that doubling the frequency does yield a statistically significant ad effect in the Full group. Intuitively, the power gains arise from approximately doubling the treatment effect since the wear-out is not too important at these exposure levels. Moreover, Mr. Wanamaker can further improve confidence by testing and imposing restrictions on the effects of ad frequency, using tests made possible by exploiting experimental variation.

Finally, we help Mr. Wanamaker find his wasted half by empirically investigating the theoretical prediction that proximity makes advertising more effective in differentiated markets. We find this to be true not only for physical proximity to the store, but also for proximity in consumer tastes, as

measured by a consumer's past loyalty to the retailer.

Ad effectiveness experiments from the field can improve our understanding of how advertising works and improve marketer's decision-making. For instance, this paper's results inspired a new product at Yahoo! called Proximity Match, which enables advertisers to target consumers who live within a specified radius of one of their stores. We anticipate that continued technological advances will reduce the cost of such field experiments (see e.g. Johnson et al. 2016b). We look forward to future research that explores the robustness of our findings and that 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.

Coey, Dominic and Michael Bailey, "People and Cookies: Imperfect Treatment Assignment in Online Experiments," in "WWW 2016" 2016.

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

Gordon, Brett R., Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky, "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," 2016. White Paper.

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

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

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

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

Johnson, Garrett A, Randall A Lewis, and David Reiley, "When Less is More: Data and Power in Advertising Experiments," *Marketing Science*, 2016. Forthcoming.

_ , _ , and Elmar I Nubbemeyer, "Ghost Ads: Improving the Economics of Measuring Ad Effectiveness," *Available at SSRN*, 2016.

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

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

Lewis, Randall A., "Measuring the Effects of Online Advertising on Human Behavior Using Natural and Field Experiments." PhD dissertation, MIT Dept of Economics 2010.

_ and Justin M. Rao, "The Unfavorable Economics of Measuring the Returns to Advertising," *Quarterly Journal of Economics*, 2015, *130* (4), 1941–1973.

Lewis, RandallA. and DavidH. 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.

Manski, C.F., *Identification for prediction and decision*, Harvard Univ Pr, 2007.

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," *Journal of Marketing Research*, 2015. Forthcoming.

_ , 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
Eligible users	1,032,204	1,032,074	1,032,299	0.988
Exposed users	572,574	571,222	570,908	0.254
Exposed Subsample				
User attributes				
Female (mean)	69.7%	69.5%	69.6%	0.216
Age (mean)	44.1	44.1	44.1	0.930
Yahoo! page views[a] (mean)	412.2	411.5	410.1	0.108
Proximity to retailer (miles, mean)	21.25	21.24	21.30	0.805
Pre-Exposure sales				
2 years prior (mean)	$848.42	$848.81	$844.67	0.546
2 weeks prior (mean)	$19.22	$19.02	$19.12	0.777
In-campaign, pre-exposure (mean)	$2.33	$2.24	$2.36	0.372
Ad exposure				
Ad views (mean)	33.42	33.41	33.66	0.349
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
Day of first exposure	3.27	3.27	3.29	0.006
Retailer ad click-through rate[b]	0.19%	0.24%	-	
Retailer ad clicker rate[c]	4.91%	3.39%	-	

Notes: Sample includes only those customers who are uniquely matched to a single Yahoo! user identifier. [a]Webpage views on Yahoo! properties during the two weeks of the campaign. [b]The click-through rate is the quotient of total ad clicks and views. [c]The clicker rate is the proportion of users exposed to the ad who click on it.

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

	(1)	(2)	(3)	(4)
Subset of Users[a]	Everyone	Exposed	Exposed	Exposed
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]Exposed 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	β [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)				
γ_θ: 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 exposed 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]	Exposed	Exposed	Exposed	Exposed	Exposed
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]Exposed 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: Histogram of Total Ad Exposures (Both Retailer and Control Ads)

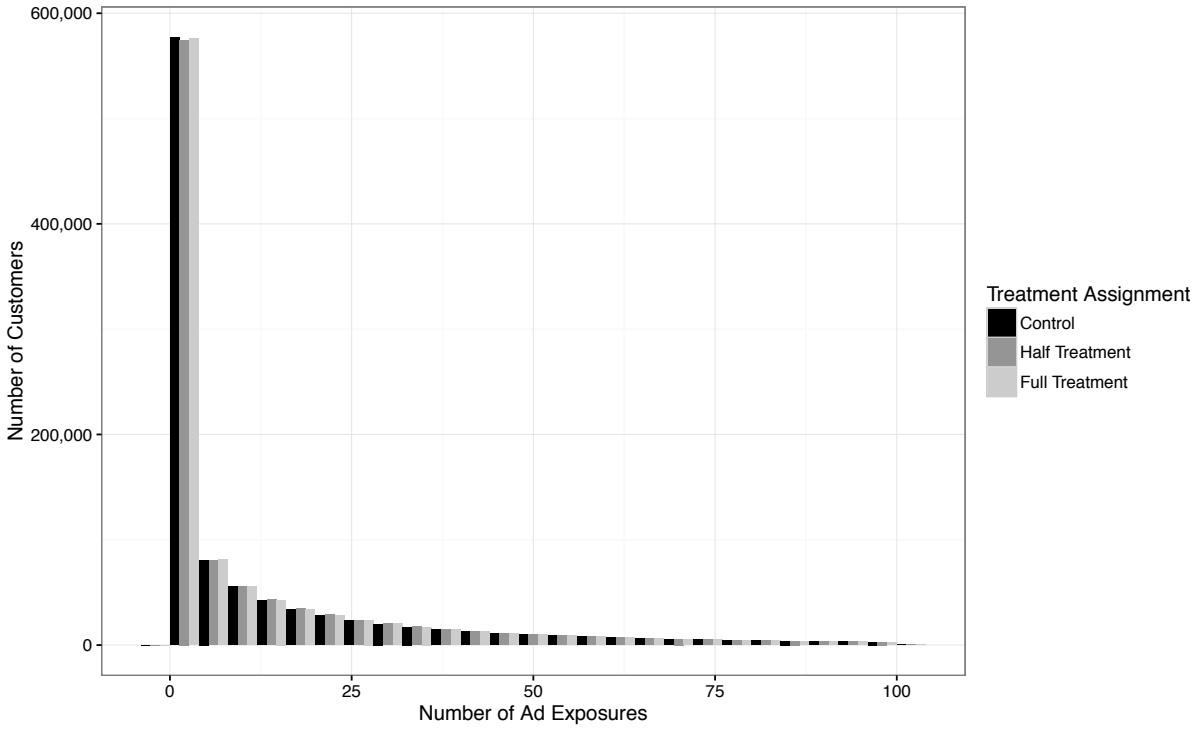


Figure 3: Histogram of Customers' Distance to the Nearest Store by Treatment Assignment

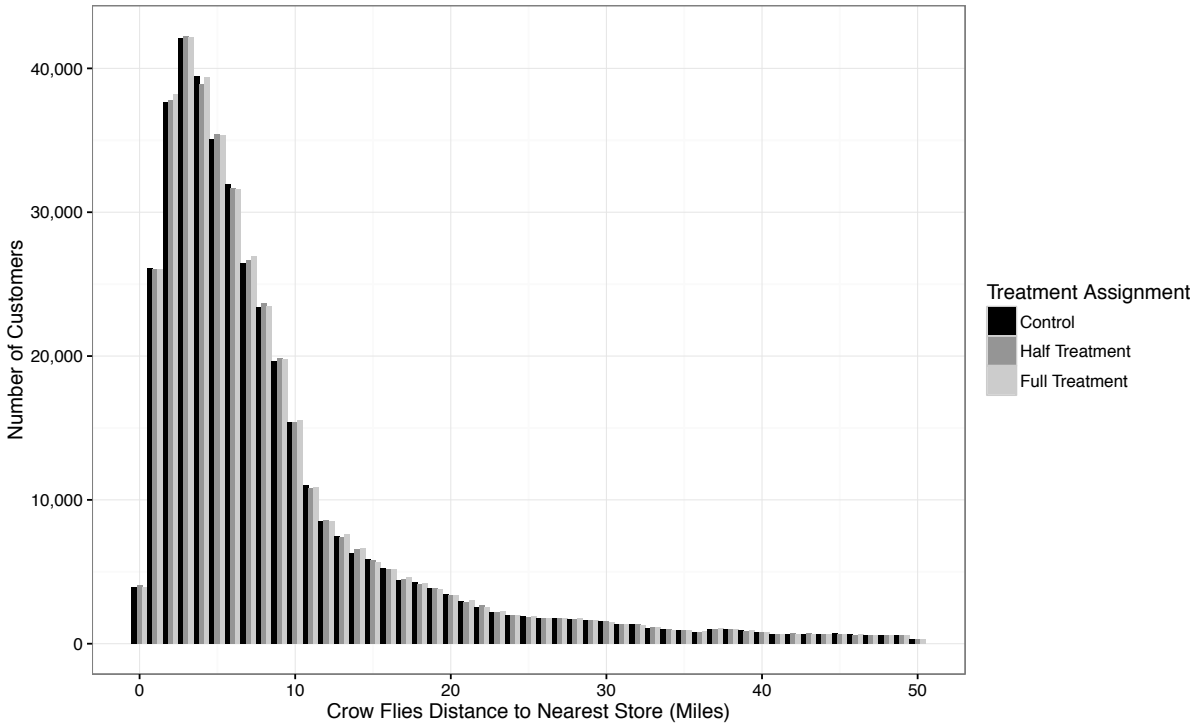


Figure 4: Joint Distribution of Retailer Ad Views (f) and Ad Type (θ)

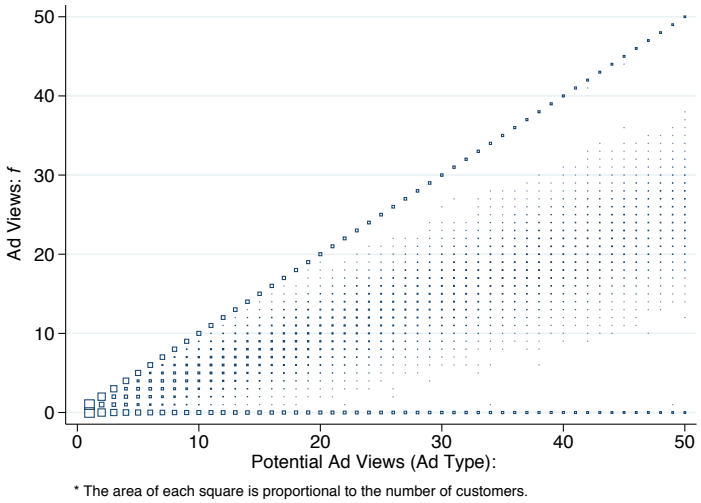


Figure 5: Indexed Sales Given a Customer's Experimental Ad Exposures (θ)

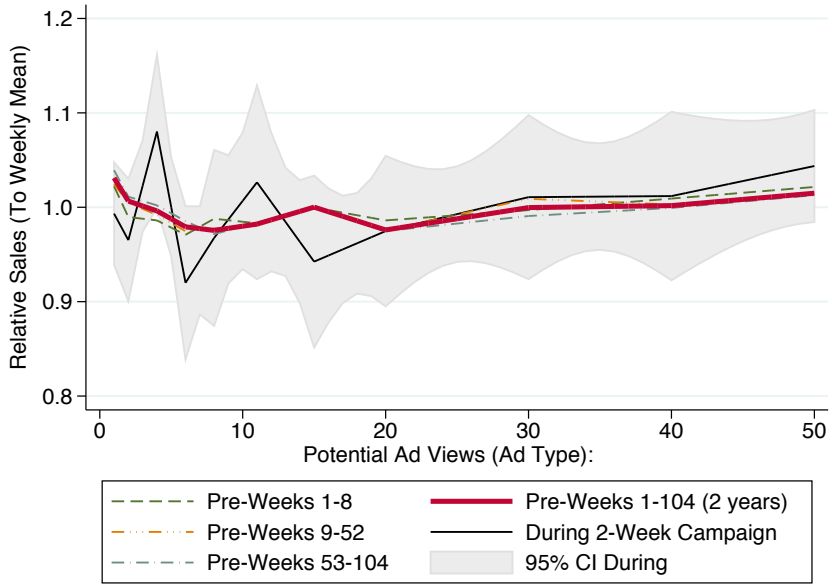


Figure 6: Frequency Response Curve: Treatment Effect for Each Value of f

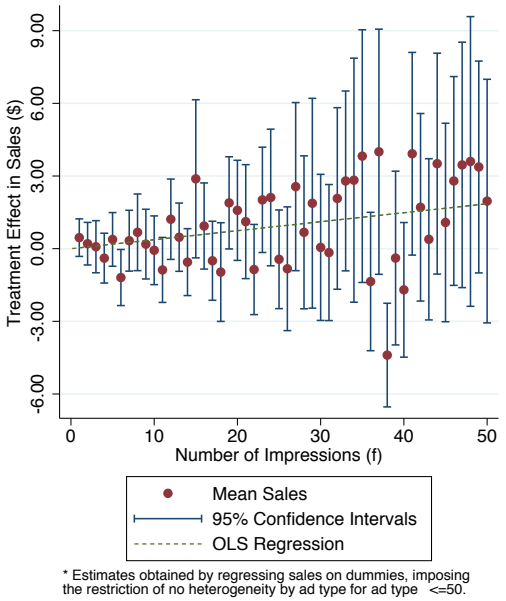
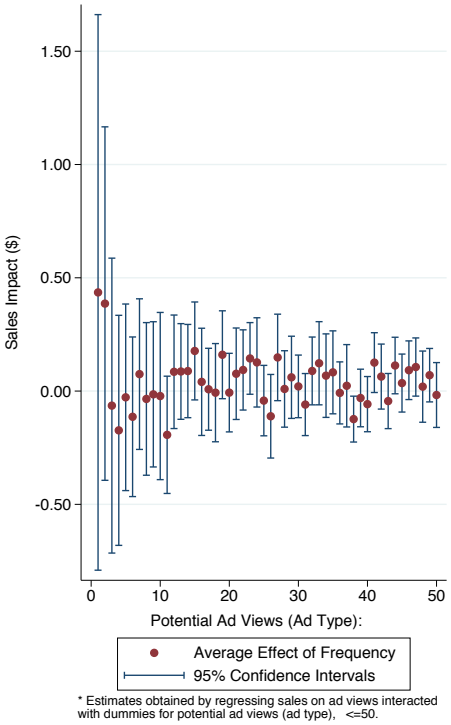


Figure 7: Marginal Impact of Ad Frequency Estimates β_θ by Ad Type (θ)



A Online Appendix: 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.

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 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)
	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
Condition					
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

Condition	(1) Earns $250,000 or more	(2) Earns $150,000 or more	(3) Earns $100,000 or more	(4) Earns $75,000 or more	(5) Earns $50,000 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	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.