Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via a Controlled Experiment on Yahoo!

Randall Lewis and David Reiley
Yahoo! Research

CCP Working Paper 11-9

Abstract: We measure the causal effects of online advertising on sales, using a randomized experiment performed in cooperation between Yahoo! and a major retailer. After identifying over one million customers matched in the databases of the retailer and Yahoo!, we randomly assign them to treatment and control groups. We analyze individual-level data on ad exposure and weekly purchases at this retailer, both online and in stores. We find statistically and economically significant impacts of the advertising on sales. The treatment effect persists for weeks after the end of an advertising campaign, and the total effect on revenues is estimated to be more than seven times the retailer's expenditure on advertising during the study. Additional results explore differences in the number of advertising impressions delivered to each individual, online and offline sales, and the effects of advertising on those who click the ads versus those who merely view them. Power calculations show that, due to the high variance of sales, our large number of observations brings us just to the frontier of being able to measure economically significant effects of advertising. We also demonstrate that without an experiment, using industry-standard methods based on endogenous crosssectional variation in advertising exposure, we would have obtained a wildly inaccurate estimate of advertising effectiveness.

ISSN 1745-9648

***Acknowledgements*:** We thank Meredith Gordon, Sergiy Matusevych, and especially Taylor Schreiner for their work on the experiment and the data. Yahoo! Incorporated provided financial and data assistance, as well as guaranteeing academic independence prior to our analysis, so that the results could be published no matter how they turned out. We acknowledge the helpful comments of Manuela Angelucci, JP Dubé, Glenn Ellison, Jerry Hausman, Kei Hirano, Garrett Johnson, John List, Preston McAfee, Sendhil Mullainathan, Paul Ruud, Michael Schwarz, Pai-Ling Yin, and participants in seminars at University of Arizona, University of California at Davis, University of California at Santa Cruz, CERGE (Prague), University of Chicago, Indian School of Business (Hyderabad), Kiev School of Economics, University of Munich, New York University, Sonoma State University, Stanford University, Vassar College, the American Economic Association meetings, the Bay Area Experimental Economics conference, the FTC Microeconomics conference, the IIOC, the Quantitative Marketing and Economics conference, and the Economic Science Association meetings in Pasadena, Lyon, and Tucson.

Contact Details:

Yahoo! Research, <ralewis@yahoo-inc.com> and <reiley@yahoo-inc.com>

Does Retail Advertising Work?

Measuring the Effects of Advertising on Sales via a Controlled Experiment on Yahoo!

Randall A. Lewis and David H. Reiley*

First Version: 21 August 2008
This Version: 8 June 2011

Abstract

We measure the causal effects of online advertising on sales, using a randomized experiment performed in cooperation between Yahoo! and a major retailer. After identifying over one million customers matched in the databases of the retailer and Yahoo!, we randomly assign them to treatment and control groups. We analyze individual-level data on ad exposure and weekly purchases at this retailer, both online and in stores. We find statistically and economically significant impacts of the advertising on sales. The treatment effect persists for weeks after the end of an advertising campaign, and the total effect on revenues is estimated to be more than seven times the retailer's expenditure on advertising during the study. Additional results explore differences in the number of advertising impressions delivered to each individual, online and offline sales, and the effects of advertising on those who click the ads versus those who merely view them. Power calculations show that, due to the high variance of sales, our large number of observations brings us just to the frontier of being able to measure economically significant effects of advertising. We also demonstrate that without an experiment, using industry-standard methods based on endogenous cross-sectional variation in advertising exposure, we would have obtained a wildly inaccurate estimate of advertising effectiveness.

* Yahoo! Research, <ralewis@yahoo-inc.com> and <reiley@yahoo-inc.com>. We thank Meredith Gordon, Sergiy Matusevych, and especially Taylor Schreiner for their work on the experiment and the data. Yahoo! Incorporated provided financial and data assistance, as well as guaranteeing academic independence prior to our analysis, so that the results could be published no matter how they turned out. We acknowledge the helpful comments of Manuela Angelucci, JP Dubé, Glenn Ellison, Jerry Hausman, Kei Hirano, Garrett Johnson, John List, Preston McAfee, Sendhil Mullainathan, Paul Ruud, Michael Schwarz, Pai-Ling Yin, and participants in seminars at University of Arizona, University of California at Davis, University of California at Santa Cruz, CERGE (Prague), University of Chicago, Indian School of Business (Hyderabad), Kiev School of Economics, University of Munich, New York University, Sonoma State University, Stanford University, Vassar College, the American Economic Association meetings, the Bay Area Experimental Economics conference, the FTC Microeconomics conference, the IIOC, the Quantitative Marketing and Economics conference, and the Economic Science Association meetings in Pasadena, Lyon, and Tucson.

I. Introduction

Measuring the causal effect of advertising on sales is a difficult problem, and very few studies have yielded clean answers. Particularly difficult has been obtaining data with exogenous variation in the level of advertising. In this paper, we present the results of a field experiment that systematically exposes some individuals but not others to online advertising, and measures the impact on individual-level sales.

With non-experimental data, one can easily draw mistaken conclusions about the impact of advertising on sales. To understand the state of the art among marketing practitioners, we consider a recent *Harvard Business Review* article (Abraham, 2008) written by the president of comScore, a key online-advertising information provider that logs the internet browsing behavior of a panel of two million users worldwide. The article, which reports large increases in sales due to online advertising, describes its methodology as follows: "Measuring the online sales impact of an online ad or a paid-search campaign—in which a company pays to have its link appear at the top of a page of search results—is straightforward: We determine who has viewed the ad, then compare online purchases made by those who have and those who have not seen it."

We caution that this straightforward technique may give spurious results. The population of people who sees a particular ad may be very different from the population who does not see an ad. For example, those people who see an ad for eTrade on the page of Google search results for the phrase "online brokerage" are a very different population from those who do not see that ad (because they did not search for that phrase). We might reasonably assume that those who search for "online brokerage" are much more likely to sign up for an eTrade account than those who do not search for "online brokerage." Thus, the observed difference in sales might not be a causal effect of ads at all, but instead might reflect a difference between these populations. In different econometric terms, the analysis omits the variable of whether someone searched for "online brokerage" or not, and because this omitted variable is correlated with sales, we get a biased estimate. (Indeed, below we will demonstrate that in our particular application, if we had used only non-experimental cross-sectional variation in advertising exposure across individuals, we would have obtained a very biased estimate of the effect of advertising on sales.) To pin down the causal effect, it would be preferable to conduct an experiment that holds the population constant between the two conditions: a treatment group of people who search for "online brokerage" would see the eTrade ad, while a control group does not see the ad.

The relationship between sales and advertising is literally a textbook example of the endogeneity problem in econometrics, as discussed by Berndt (1991) in his applied-econometrics text. Theoretical work by authors such as Dorfman and Steiner (1954) and Schmalensee (1972) shows that we might expect advertisers to choose the optimal level of advertising as a function of sales, so that regressions to determine advertising's effects on sales are plagued by the possibility of reverse causality. Berndt (1991) reviews a substantial econometric literature on this topic.

After multiple years of interactions with advertisers and advertising sales representatives at Yahoo!, we have noticed a distinct lack of knowledge about the quantitative effects of advertising. This suggests that the economic theory of advertising has likely gotten ahead of practice, in the sense that advertisers (like Wanamaker) typically do not have enough quantitative information to be able to choose optimal levels of advertising. They may well choose advertising budgets as a fraction of sales (producing econometric endogeneity, as discussed in Berndt (1991)), but these are likely rules of thumb rather than informed, optimal decisions. Systematic experiments, which might measure the causal effects of advertising, are quite rare in practice.

Most advertisers do not systematically vary their levels of advertising to measure the effects on sales. Notable exceptions include direct-mail advertising, where advertisers do run frequent experiments (on advertising copy, targeting techniques, etc.) in order to measure direct-response effects by consumers. In this study, we address brand advertising, where the expected effects have to do with longer-term consumer goodwill rather than direct responses. In this field, advertising's effects are much less well understood. Advertisers often change their levels of advertising over time, as they run discrete "campaigns" during different calendar periods, but this variation does not produce clean data for measuring the effects of advertising because other variables also change concurrently over time. For example, if a retailer advertises more during December than in other months, we do not know how much of the increased sales to attribute to the advertising, and how much to increased holiday demand.

As is well known in the natural sciences, experiments are a great way to establish and measure causal relationships. Randomizing a policy across treatment and control groups allows us to vary advertising in a way that is uncorrelated with all other factors affecting sales, thus eliminating econometric problems of endogeneity and omitted-variable bias. This recognition has become increasingly important in economics and the social science; see Levitt and List

(2008) for a summary. We add to this recent literature with an unusually large-scale field experiment involving over one million subjects.

A few previous research papers have also attempted to quantify the effects of advertising on sales through field experiments. Several studies have made use of IRI's BehaviorScan technology, a pioneering technique developed for advertisers to experiment with television ads and measure the effects on sales. These studies developed panels of households whose sales were tracked with scanner data and split the cable-TV signal to give increased exposures of a given television ad to the treatment group relative to the control group. The typical experimental sample size was approximately 3,000 households. Abraham and Lodish (1990) report on 360 studies done for different brands, but many of the tests turned out to be statistically insignificant. Lodish *et al.* (1995a) report that only 49% of the 360 tests were significant at the 20% level (one-sided), and then go on to perform a meta-analysis showing that much of the conventional wisdom among advertising executives did not help to explain which ads were relatively more effective in influencing sales. Lodish *et al.* (1995b) investigated long-run effects, showing that for those ads that did produce statistically significant results during a year-long experiment, there tended to be positive effects in the two following years as well. Hu, Lodish, and Krieger (2007) perform a follow-up study and find that similar tests conducted after 1995 produce larger impacts on sales, though more than two thirds of the tests remain statistically insignificant.

The lack of statistical significance in these previous experimental tests likely reflects low statistical power. As we shall show in this paper, an economically significant effect of advertising (one that generates a positive return on the cost of the ads) could easily fail to be statistically significant even in a clean experiment with hundreds of thousands of observations per treatment. The variance of sales can be quite high, and an advertising campaign can be economically profitable even when it explains only a tiny fraction of sales. Looking for the effects of brand advertising can therefore resemble looking for a needle in a haystack. By studying over a million users, we are finally able to shrink confidence intervals to the point where effects of economically interesting magnitudes have a reasonable chance of being statistically significant.

More recently, Anderson and Simester (2008) experimented with a catalog retailer's frequency of catalog mailings, a direct-mail form of retail advertising. A sample of 20,000 customers received either twelve or seventeen catalog mailings over an eight-month period.

When customers received more mailings, they exhibited increased short-run purchases. However, they also found evidence of intertemporal substitution, with the firm's best customers making up for short-run increases in purchases with longer-run decreases in purchases.

Ackerberg (2001, 2003) makes use of non-experimental individual-level data on yogurt advertising and purchases for 2000 households. By exploiting the panel nature of the dataset, he shows positive effects of advertising for a new product (Yoplait 150), particularly for consumers previously inexperienced with the product. For a comprehensive summary of theoretical and empirical literature on advertising, see Bagwell (2005).

Because our data, like Ackerberg's, has a panel structure with individual sales data both before and after the advertising campaign, we employ a difference-in-difference (DID) estimator that exploits both experimental and non-experimental variation in advertising exposure. The DID estimator yields a very similar point estimate to the simple experimental difference, but with higher precision. We therefore prefer the more efficient DID estimate, despite the need to impose an extra identifying assumption (any time-varying individual heterogeneity in purchasing behavior must be uncorrelated with advertising exposure). Though our preferred estimator could in principle have been computed on purely observational data, we still rely heavily on the experiment for two reasons: (1) the simple experimental difference tests the DID identifying assumption and makes us much more confident in the results than would have been possible with standard observational data, and (2) the experiment generates substantial additional variance in advertising exposure, thus increasing the efficiency of the estimate.

The remainder of this paper is organized as follows. We present the design of the experiment in Section II, followed by a description of the data in Section III. In Section IV, we measure the effect on sales during the first of two[1] advertising campaigns in this experiment. In Section V, we demonstrate and measure the persistence of this effect after the campaigns have ended. In Section VI we return to the first campaign, the larger and more impactful of the two we conducted, to examine how the treatment effect of online advertising varies across a number of dimensions. This includes the effect on online versus offline sales, the effect on those who click ads versus those who merely view them, the effect for users who see a low versus high frequency

[1] Previous drafts of this paper examined three campaigns, but specification tests called into question the reliability of the difference-in-differences estimator applied to the mismatched merge required to combine the third campaign's sales data with the first two campaigns. The first two campaigns were already joined via a unique identifier unavailable in the third campaign's data. We now omit all references to the third campaign for reasons of data reliability and simplicity.

of ads, and the effect on number of customers purchasing versus the size of the average purchase. The final section concludes.

II. Experimental Design

This experiment randomized individual-level exposure to a nationwide retailer's display-advertising campaign on Yahoo! This enabled us to measure the causal effects of the advertising on individuals' weekly purchases, both online and in stores. To achieve this end, we matched the retailer's customer database against Yahoo!'s user database. This match yielded a sample of 1,577,256 individuals who matched on name and either email or postal address. Note that the population under study is therefore the set of existing customers of the retailer who log in to Yahoo! [2]

Of these matched users, we assigned 81% to a treatment group who subsequently viewed two advertising campaigns on Yahoo! from the retailer. The remaining 19% were assigned to the control group and saw none of the retailer's ads on Yahoo! The simple randomization was designed to make the treatment-control assignment independent of all other relevant variables.

The treatment group of 1.3 million Yahoo! users was exposed to two different advertising campaigns over the course of two months in fall 2007, separated by approximately one month. Table 1 gives summary statistics for the campaigns, which delivered 32 million and 10 million impressions, respectively. By the end of the second campaign, a total of 868,000 users had been exposed to ads. These individuals viewed an average of 48 ad impressions per person.

These represent the only ads shown by this retailer on Yahoo! during this time period. However, Yahoo! ads represent a small fraction of the retailer's overall advertising budget, which included other media such as newspaper and direct mail. As we shall see, Yahoo! advertising explains a very small fraction of the variance in weekly sales. But because of the randomization, the Yahoo! advertising is uncorrelated with any other influences on shopping behavior, and therefore our experiment gives us an unbiased estimate of the causal effects of the advertising on sales.

The campaigns in this experiment consisted of "run-of-network" ads on Yahoo! This means that ads appeared on various Yahoo! properties, such as mail.yahoo.com,

[2] The retailer gave us some portion of their entire database, probably selecting a set of customers they were most interested in advertising to. We do not have precise information about their exact selection rule.

groups.yahoo.com, and maps.yahoo.com. Figure 1 shows a typical display advertisement placed on Yahoo! The large rectangular ad for Netflix[3] is similar in size and shape to the advertisements in this experiment.

Following the experiment, Yahoo! and the retailer sent data to a third party who matched the retail sales data to the Yahoo! browsing data. The third party then anonymized the data to protect the privacy of customers. In addition, the retailer disguised actual sales amounts by multiplying by an undisclosed number between 0.1 and 10. Hence, all financial quantities involving treatment effects and sales will be reported in R$, or "Retail Dollars," rather than actual US dollars.

III.Sales and Advertising Data

Table 2 provides some summary statistics for the first campaign, providing evidence consistent with a valid randomization.[4] The treatment group was 59.7% female while the control group was 59.5% female, a statistically insignificant difference (p=0.212). The proportion of individuals who did any browsing on the Yahoo! network during the campaign was 76.4% in each group (p=0.537). Even though 76.4% of the treatment group visited Yahoo! during the campaign, only 63.7% of the treatment group actually received pages containing the retailer's ads. On average, a visitor received the ads on only 7.0% of the pages she visited. The probability of being shown an ad on a particular page depends on a number of variables, including user demographics, the user's past browsing history, and the topic of the page visited.

The number of ads viewed by each Yahoo! user in this campaign is quite skewed. The very large numbers in the upper tail are likely due to the activity of non-human "bots," or automated browsing programs. Restricting attention to users in the retail database match should tend to reduce the number of bots in the sample, since each user in our sample has previously made a purchase at the retailer. Nevertheless, we still see a small number of likely bots, with

[3] Netflix was not the retailer featured in this campaign but is an example of a firm which only does sales online and advertises on Yahoo! The major retailer with whom we ran the experiment prefers to remain anonymous.

[4] Only one statistic in this table is statistically significantly different across treatment groups. The mean number of Yahoo! page views was 363 pages for the treatment group versus 358 for the control group, a statistically but not economically significant difference (p=0.0016). The significant difference comes largely from the outliers at the top of the distribution, as almost all of the top 30 page viewers ended up being assigned to the treatment group. If we trim the top 250 out of 1.6 million individuals from the dataset (that is, removing all the bot-like individuals with 12,000 or more page views in two weeks), the difference is no longer significant at the 5% level. The lack of significance remains true whether we trim the top 500, 1000, or 5000 observations from the data.

extreme browsing behavior. Figure 2 shows a frequency histogram of the number of the retailer's ads viewed by treatment group members that saw at least one of the ads during campaign #1. The majority of users saw fewer than 100 ads, with a mere 1.0% viewing more than 500 ads during the two weeks of the online ad campaign. The maximum number of the ads delivered to a single individual during the campaign was 6050.[5]

One standard statistic in online advertising is the click-through rate, or fraction of ads that were clicked by a user. The click-through rate for this campaign was 0.28%. With detailed user data, we can also tell that, conditional on receiving at least one ad, the proportion of the designated treatment group who clicked at least one ad in this campaign was 7.2% (sometimes called the "clicker rate").

In order to protect the privacy of individual users, a third party matched the retailer's sales data to the Yahoo! browsing data and anonymized all observations so that neither party could identify individual users in the matched dataset. This weekly sales data includes both online and offline sales and spans approximately 18 weeks: 3 weeks preceding, 2 weeks during, and 1 week following each of the two campaigns. Sales amounts include all purchases that the retailer could link to each individual customer in the database.[6]

Table 3 provides a weekly summary of the sales data, while Figure 3 decomposes the sales data into online and offline components. We see that offline (in-store) sales represent 86% of the total. Combined weekly sales are quite volatile, even though averaged across 1.6 million individuals, ranging from less than R$0.60 to more than R$1.60 per person. The standard deviation of sales across individuals is much larger than the mean, at approximately R$14. The mean includes a large mass of zeroes, as fewer than 5% of individuals in a given week make any transaction (see last column of Table 3). For those who do make a purchase, the transaction amounts exhibit large positive and negative amounts, but well over 90% of purchase amounts lie between –R$100 and +R$200. Negative purchase amounts represent net returns of merchandise; we do not exclude these observations from our analysis because advertising could easily cause a customer's total purchases in a week to be less negative than they would otherwise.

[5] Although the data suggests extreme numbers of ads, Yahoo! engages in extensive anti-fraud efforts to ensure fair pricing of its products and services. In particular, not all ad impressions in the dataset were deemed valid impressions and charged to the retailer.

[6] To the extent that these customers make purchases that cannot be tracked by the retailer, our estimate may underestimate the total effect of advertising on sales. However, the retailer believes that it correctly attributes 90% of purchases to the correct individual customer. They use several methods to attribute purchases to the correct customer account, such as matching the name on a customer's credit card at checkout.

The high variance in the data implies surprisingly low power for our statistical tests. Many economists have the intuition that a million individual observations is approximately infinite, meaning that any economically interesting effect of advertising must be highly statistically significant. This intuition turns out to be incorrect in our setting, where the variance of individual purchases (driven by myriad idiosyncratic factors) makes for a rather large haystack in which to seek the needle of advertising's effects.

For concreteness, suppose hypothetically that our first advertising campaign were so successful that the firm obtained a 100% short-run return on its investment. The campaign cost approximately R$25,000 to the retailer[7], representing R$0.02 per member of the treatment group, so a 100% return would represent a R$0.04 increase in cash flow due to the ads. Consultation with retail-industry experts leads us to estimate this retailer's margins to be approximately 50% (if anything, we have estimated this to be conservatively low). Then a cash-flow increase of R$0.04 represents incremental revenues of R$0.08, evenly divided between the retail margin and the cost of goods sold. These hypothesized incremental revenues of R$0.08 represent a 4% increase in the mean sales per person (R$1.89) during the two weeks of the campaign. With such a successful advertising campaign, how easy would it be to reject the null hypothesis of no effect of advertising?

Note that the standard deviation of two-week sales (R$19) is approximately ten times the mean level of sales, and 250 times the size of the true treatment effect. Thus, even with over 300,000 control-group members and 1,200,000 treatment-group members, the standard deviation of the difference in sample means will remain as large as R$0.035. This gives confidence intervals with a width of ±R$0.07 when we hope to detect an effect of R$0.08. Under our specified alternative hypothesis of the retailer doubling its money, the probability of finding a statistically significant effect of advertising with a two-tailed 5% test is only 63%. With a smaller hypothesized increase in advertising revenues – assume the retailer only breaks even on its advertising dollars with a revenue increase of only R$0.04 - the probability of rejection is only 21%. These power calculations demonstrate surprisingly high probability of type-II error, indicating that the very large scale of our experiment puts us exactly at the measurement frontier

[7] Because of the custom targeting to the selected database of known retailer customers, Yahoo! charged the retailer an appropriately higher rate, on the order of five times the price that would normally be charged for an equivalent untargeted campaign. In our return-on-investment calculations, we use the actual price (for custom targeting) charged to the retailer.

where we can hope to detect statistically significant effects of an economically meaningful campaign.[8]

In our data description in this section, we have focused mainly on the first of the two campaigns in our experiment. We have done this for two reasons. First, the first campaign accounts for more than 75% of the total number of ad impressions, so we expect its effects to be much larger. Second, both campaigns were shown to the same treatment and control groups, which prevents us from estimating the separate effects of campaign #2 if advertising has persistent effects across weeks. In section V, we will present evidence of such persistence and give a combined estimate of the combined effects of campaigns #1 and #2. For simplicity, we begin with estimating the isolated effects of the larger and earlier of the two campaigns.

IV. Basic Treatment Effect in Campaign #1

For campaign #1 we are primarily interested in estimating the effect of the treatment on the treated individuals. In traditional media such as TV commercials, billboards, and newspaper ads, the advertiser must pay for the advertising space, regardless of the number of people that actually see the ad. With online display advertising, by contrast, it is a simple matter to track potential customers and standard to bill an advertiser by the number of delivered ad impressions. While there is an important difference between a delivered ad and a seen ad, our ability to count the number of attempted exposures gives us fine-grained ability to measure the effects of the impressions paid for by the advertiser.

Table 4 gives initial results comparing sales between treatment and control groups. We look at total sales (online and offline) during the two weeks of the campaign, as well as total sales during the two weeks prior to the start of the campaign.[9] During the campaign, we see that the treatment group purchased R$1.89 per person, compared to the control group at $1.84 per person. This difference gives a positive estimate of the effect of the treatment effect on the intent

[8] This back-of-the envelope analysis helps us understand why Lodish et al. (1995a) used a 20% one-sided test as their threshold for statistical significance, a level that at first seemed surprisingly high to us, relative to conventional hypothesis tests. This is especially true when we remember that their sample sizes were closer to 3,000 than to our 1.6 million.

[9] Though we have three weeks of pre-period data available, we have chosen to use only two weeks here, for reasons of symmetry and simplicity of exposition (two weeks are intuitively comparable to two weeks). In order to see the same results using a three-week pre-period baseline, please see Section V, particularly Table 5.

to treat with ads of R$0.053 (0.038) per person. The effect is not statistically significant at the 5% level (p=0.162, two-sided).[10]

For the two weeks before the campaign, the control group purchased slightly (and statistically insignificantly) more than the treatment group: R$1.95 versus R$1.93. We can combine the pre- and post-campaign data to obtain a difference-in-difference estimate of the increase in sales for the treatment group relative to the control (again, an estimate of the effect of the intent to treat). This technique gives a slightly larger estimate of R$0.064 per person, but is again statistically insignificant at conventional levels (p=0.227).

Because only 64% of the treatment group was actually treated with ads, this simple treatment-control comparison has been diluted with the 36% of individuals who did not see any ads during this campaign (due to their individual browsing behavior). Since the advertiser pays per impression, they care only about the effect of advertising on those individuals who actually received ads. Ideally, we would remove the unexposed 36% of individuals both from the treatment and control groups in order to get an estimate of the treatment effect on the treated. Unfortunately, we are unable to observe which control-group members would have seen ads for this campaign had they been in the treatment group,[11] so we cannot remove the statistical noise of the endogenously untreated individuals. However, we can at least compute an unbiased estimate of the treatment effect on the treated. We scale up our diluted treatment effect (R$0.05) by dividing by 0.64, the fraction of individuals treated,[12] for an estimate of the treatment effect

[10] However, it does easily exceed the significance threshold used to assess a campaign as successful in Lodish *et al.* (1995a).

[11] We recorded zero impressions of the retail ad campaign to every member of the control group, which makes it impossible to distinguish those control group members who would have seen ads. The Yahoo! ad server uses a complicated set of rules and constraints to determine which ad will be seen by a given individual on a given page. For example, a given ad might be shown more often on Yahoo! Mail than on Yahoo! Finance. If another advertiser has targeted females under 30 during the same time period, then this ad campaign may have been relatively more likely to be seen by other demographic groups. Our treatment-control assignment represented an additional constraint. Because of the complexity of the server delivery algorithm, we were unable to model the hypothetical distribution of ads delivered to the control group with an acceptable level of accuracy. Therefore, we cannot restrict attention to treated individuals without risking considerable selection bias in our estimate of the treatment effect.

[12] This is equivalent to estimating the local average treatment effect (LATE) via instrumental variables via the following model:

$$Sales_{i,t} = \gamma_t SawAds_{i,t} + \beta_t + \varepsilon_{i,t}$$
$$SawAds_{i,t} = \pi_{0,t} + \pi_{1,t} Treatment_i + \eta_{i,t}$$

where the first stage regression is an indicator for whether the number of the retailer's ads seen is greater than zero on the exogenous treatment-control randomization. As such, this transforms our intent-to-treat estimates into estimates of the treatment on the treated.

on those treated with ads: R$0.083 (0.059). The standard error is also scaled proportionally, leaving the level of statistical significance unaffected (p=0.162).

Now suppose that instead of running an experiment, we had instead estimated the effects of advertising by a cross-sectional observational study, as in Abraham (2008). We would not have an experimental control group, but would instead be comparing the endogenously treated versus untreated individuals. We can see from the last two lines of Table 4 that instead of an increase of R$0.083 due to ads, we would instead have estimated the difference to be –R$0.23! The difference between the exposed consumers (R$1.81) and the unexposed consumers (R$2.04) is opposite in sign to the true estimated effect, and would have been reported as highly statistically significant. This allows us to quantify the selection bias that would result from a cross-sectional comparison of observational data: R$0.31 lower than the unbiased experimental estimate of R$0.083.

This selection bias results from heterogeneity in shopping behavior that happens to be correlated with ad views: in this population, those who browse Yahoo! more actively also have a tendency to purchase less at the retailer, independent of ad exposure. We see this very clearly in the pre-campaign data, where those treatment-group members who would eventually see online ads purchased considerably less (R$1.81) than those who would see no ads (R$2.15). This statistically significant difference ($p<0.01$) confirms our story of heterogeneous shopping behavior negatively correlated with Yahoo! browsing and ad delivery, and the large bias that can result from cross-sectional study of advertising exposure in the absence of an experiment.

For our present purposes, we see that it would be a mistake to exclude from the study those treatment-group members who saw no online ads, because the remaining treatment-group members would not represent the same population as the control group. Such an analysis would result in selection bias towards finding a negative effect of ads on sales, because the selected treatment-group members purchase an average of R$1.81 in the absence of any advertising treatment, while the control-group members purchase an average of R$1.95—a statistically significant difference of R$0.13 (p=0.002).

During the campaign, there persists a sales difference between treated and untreated members of the treatment group, but this difference becomes smaller. While untreated individuals' sales drop by R$0.10 from before the campaign, treated individuals' sales remained constant. Furthermore, control-group mean sales also fell by R$0.10 during the same period, just

like the untreated portion of the treatment group. This suggests that advertisements may be preventing treated individuals' sales from falling like untreated individuals' sales did. This will lead us to our preferred estimator below, a difference in differences between treated and untreated individuals (where "untreated" pools together both control-group members and untreated members of the designated treatment group).

Before presenting our preferred estimator, we first look at the shape of the distribution of sales. Figure 4 compares histograms of sales amounts for the treatment group and control group, omitting those individuals for whom there was no transaction. For readability, these histograms exclude the most extreme outliers, trimming approximately 0.5% of the positive purchases from both the left and the right of the graph.[13] Relative to the control, the treatment density has less mass in the negative part of the distribution (net returns) and more mass in the positive part of the distribution. These small but noticeable differences both point in the direction of a positive treatment effect, especially when we recall that this diagram is diluted by the 34% of customers who did not browse enough to see any ads on Yahoo! Figure 5 plots the difference between the two histograms in Figure 4. The treatment effect is the average over this difference between treatment and control sales distributions.

Next we exploit the panel nature of our data by using a difference-in-differences (DID) model. This allows us to estimate the effects of advertising on sales while controlling for the heterogeneity we have observed across individuals in their purchasing behavior. Our DID model makes use of the fact that we observe the same individuals both before and after the start of the ad campaign. We begin with the following model:

$$Sales_{i,t} = \gamma_t SawAds_{i,t} + \beta_t + \alpha_i + \varepsilon_{i,t}.$$

In this equation, $Sales_{i,t}$ is the sales for individual i in time period t, $SawAds_{i,t}$ is the dummy variable indicating whether individual i saw any of the retailer's ads in time period t, γ_t is the average effect of viewing the ads, β_t is a time-specific mean, α_i is an individual effect or unobserved heterogeneity (which we know happens to be correlated with viewing ads), and $\varepsilon_{i,t}$ is

[13] We trim about 400 observations from the left and 400 observations from the right from a total of 75,000 observations with nonzero purchase amounts. These outliers do not seem to be much different between treatment and control. We leave all outliers in our analysis, despite the fact that they increase the variance of our estimates. Because all data were recorded electronically, we have no reason to suspect coding errors.

an idiosyncratic disturbance. Computing time-series differences will enable us to eliminate the individual unobserved heterogeneity α_i.

We consider two time periods: (1) the "pre" period of two weeks before the start of campaign #1, and (2) the "post" period of two weeks after the start of the campaign. By computing first differences of the above model across time, we obtain:

$$Sales_{i,post} - Sales_{i,pre} = \gamma_t SawAds_{i,post} - \gamma_t SawAds_{i,pre} + \beta_{post} - \beta_{pre} + \varepsilon_{i,post} - \varepsilon_{i,pre}$$

Since no one saw ads in the "pre" period, we know that $SawAds_{i,pre} = 0$. So the difference equation simplifies to:

$$\Delta Sales_i = \gamma_t SawAds_{i,post} + \Delta\beta + \Delta\varepsilon_i$$

We can then estimate this difference equation via ordinary least squares (OLS). The gamma coefficient is directly comparable to the previous (rescaled) estimate of R$0.083 (0.059) of the effect of the treatment on the treated. Note that in this specification, unlike the previous specifications, we pool together everyone who saw no ads in the campaign, including both the control group and those treatment-group members who turned out not to see any ads.

Using difference in differences, the estimated average treatment effect of being treated by viewing at least one of the retailer's ads during the campaign is R$0.102, with a standard error of R$0.043. This effect is statistically significant (p=0.018) as well as economically significant, representing an average increase of 5% on treated individuals' sales. Based on the 814,052 treated individuals, the estimate implies an increase in revenues for the retailer of R$83,000 ± 68,000 (95% confidence interval) due to the campaign. Because the cost of campaign #1 was approximately R$25,000,[14] the point estimate suggests that the ads produced more than 325% as much revenue as they cost the retailer. We conclude that retail advertising does, in fact, work.

The main identifying assumption of the DID model is that each individual's idiosyncratic tendency to purchase from the retailer is constant across time, and thus the treatment variable is

[14] These advertisements were more expensive than a regular run-of-network campaign. The database match was a form of targeting that commanded a large premium. In our cost estimates, we report the dollar amounts (scaled by the retailer's "exchange rate") actually paid by the retailer to Yahoo!

uncorrelated with the DID error term. That is, while individual purchase levels are correlated with ad exposure, we assume that individual time-series differences are not. This assumption could be violated if some external event, either before or during the experiment, had different effects on the retail purchases of those who did versus did not see ads. For example, perhaps in the middle of the time period studied, the retailer did a direct-mail campaign we do not know about, and the direct mail was more likely to reach those individuals in our study who browsed less often on Yahoo! Fortunately, our previous, experimentally founded, estimates are very similar in magnitude to the DID estimates: R$0.083 for the simple comparison of levels between treatment and control, versus R$0.102 for the DID estimate.

Note that even when we are exploiting non-experimental variation in the data, we still make important use of the fact that we ran an experiment. The similarity between these two estimates reassures us about the validity of our DID specification. We note that there are two distinctions between our preferred DID estimate and our original treatment-control estimate. First, DID looks at pre-post differences for each individual. Second, DID compares between treated and untreated individuals (pooling part of the treatment group with the control group), rather than simply comparing between treatment and control groups. We perform a formal specification test of this latter difference by comparing pre-post sales differences in the control group versus the untreated portion of the treatment group. The untreated portion of the treatment group has a mean just R$0.001 less than the mean of the control group, and we cannot reject the hypothesis that these two groups are the same (p=0.988).

Finally, we note that if there is any measurement error of customer assignment to treatment versus control, then our estimates will underestimate the effects of the advertising. There are several ways that mismatching of sales and advertising data could occur. For example, the third party who matched the data could have allowed for imperfect matches, such as assuming that the Barbara Smith who lives on Third Street is the same as the Barbara Smith who lives on Fifth Street. (We do not know the exact algorithm used by the third party.) Another example is that if a husband is browsing Yahoo while his wife is logged in to the home computer, we might assume she was exposed to the advertising though in fact she was not. In either of these cases, we could be measuring someone's sales assuming they were treated even though we never delivered ads to them. We would therefore measure a smaller treatment difference than the actual truth, through attenuation bias.

V. Persistence of the Effects

Our next question concerns the longer-term effects of the advertising after the campaign has ended. One possible case is that the effects could be persistent and increase sales even after the campaign is over. Another case is that the effects are short-lived and only increase sales during the period of the campaign. A third possibility is that advertising could have negative long-run effects if it causes intertemporal substitution by shoppers: purchasing today something that they would otherwise have purchased a few weeks later. In this section, we distinguish empirically between these three competing hypotheses.

A. Sales in the Week after the Campaign Ended

We begin by focusing on the six weeks of data which we received from the retailer tailored to the purposes of analyzing campaign #1. As previously mentioned, this includes three weeks of data prior to campaign #1 and three weeks following its start. To perform the test of the above hypotheses, we use the same Difference-in-Differences model as before, but this time include in the "post" period the third week of sales results following the start of two-week campaign. For symmetry, we also use all three weeks of sales in the "pre" period, in contrast to the results in the previous section, which were based on two weeks both pre and post. As before, the DID model compares the pre-post difference for treated individuals with the pre-post difference for untreated individuals (including both control-group members and untreated treatment-group members).

Before presenting our estimate, we first show histograms in Figure 6 of the distributions of three-week pre-post sales differences. Note three differences between Figure 6 and the histogram presented earlier in Figure 4: (1) we compare pre-post differences rather than levels of sales, (2) we compare treated versus untreated individuals rather than treatment versus control groups, and (3) we look at three weeks of sales data (both pre and post) rather than just two. The difference between the treated and untreated histograms can be found in Figure 7, with 95% confidence intervals for each bin indicated by the whiskers on each histogram bar. We see that the treated group has substantially more weight in positive sales differences and substantially less weight in negative sales differences. This suggests a positive treatment effect, which we now proceed to measure via difference in differences. Using our preferred DID estimator, we find that the estimated treatment effect increases from R$0.102 for two weeks to R$0.166 for three weeks.

Thus, the treatment effect for the third week appears to be just as large as the average effect per week during the two weeks of the campaign itself. To pin down the effects in the third week alone, we run a DID specification comparing the third week's sales with the average of the three pre-campaign weeks' sales. This gives us an estimate of R$0.061 with a standard error of R$0.024 (p=0.01), indicating that the effect in the third week is both statistically and economically significant. Importantly, the effect in the week after the campaign (R$0.061) is just as large as the average per-week effect during the two weeks of the campaign (R$0.051).

B. More than One Week after the Campaign Ended

Could the effects be persistent even more than a week after the campaign ends? We investigate this question using sales data collected for purposes of evaluating campaign #2. Recall that for both campaigns, we obtained three weeks of sales data before the start of the campaign, and three weeks of sales data after the start of the campaign. It turns out that the earliest week of pre-campaign sales for campaign #2 happens to be the fourth week after the start of campaign #1, so we can use that data to examine the treatment effect of campaign #1 in its fourth week.[15]

In order to check for extended persistence of advertising, we use the same DID model as before, estimated on weekly sales. Our "pre-period" sales will be the weekly average of sales in the three weeks preceding the start of campaign #1. Our "post-period" sales will be the sales during a given week after the start of campaign #1. We then compute a separate DID estimate for each week, beginning with the first week of campaign #1 and ending with the week following campaign #2.[16]

Table 5 displays the results, and Figure 8 represents them graphically. In the figure, vertical lines indicate the beginning (solid) and end (dashed) of both campaigns. The estimated

[15] Because the campaigns did not start and end on the same day of the week, we end up with a three-day overlap between the third week after the start of campaign #1 and the third week prior to the start of campaign #2. That is, those three days of sales are counted twice. We correct for this double-counting in our final estimates of the total effect of advertising on sales by scaling the estimates by the ratio of the number of weeks the data spans to the number of weeks the data fields represent. In aggregate estimates over the entire period, this is the ratio of 8 weeks to 8 weeks and 3 days, due to the 3-day double-counting.

[16] Because campaign #2 lasted ten days rather than an even number of weeks, the second "week" of the campaign consists of only three days instead of seven. In this case of a 3-day "week," we scale up the sales data by 7/3 to keep consistent units of sales per week. This implicitly assumes that purchasing behavior and treatment effects are the same across days of the week, which seems implausible to us, but we are unconcerned because that three-day "week" represents a relatively minor part of the overall analysis.

treatment effects in later weeks thus include cumulative effects of the campaigns run to date. The average weekly treatment effect on the treated is R$0.036, with individual weekly estimates ranging from R$0.004 to R$0.061. Although most of the individual weekly treatment effects are statistically indistinguishable from zero (95% confidence intervals graphed in Figure 8), we find it striking that every single point estimate is positive.[17] We particularly note the large, positive effects estimated during the inter-campaign period, more than three weeks after ads stopped showing for the retailer's first campaign on Yahoo!

To obtain an estimate of the cumulative effect of both campaigns, we use all nine weeks of data. For econometric efficiency, we compute an average of the nine weekly estimates of the treatment effect, taking care to report standard errors that account for the covariances between regression coefficients across weeks. Table 6 reports an optimally weighted average of the nine per-week treatment effects,[18] with a simple average included for comparison. The weighted average is R$0.039 (R$0.0147) per week. We then multiply this number by eight to get a total effect across the entire time period of observation, since the "nine-week" time period actually includes a total of only eight weeks. [19] This multiplication gives us R$0.311 (R$0.117).

To estimate the total benefit of the two campaigns, we take our estimate of R$0.311 and multiply it by the average number of users who had already been treated with ads in a given week, which turns out to be 812,000. This gives us a 95% confidence interval estimate of the total incremental revenues due to ads of R$253,000 ± 188,000. For comparison, the total cost of

---

[17] To avoid overstating the significance of these observation, we note that the weekly estimates are not independent of each other. Each week's DID estimator relies on the same three weeks of pre-campaign data.

[18] We implement the weighted average by computing a standard GLS regression on a constant, where the GLS weighting matrix is the covariance matrix among the nine regression coefficients. These covariances can be analytically computed for two different weeks, *j* and *k*, as

$$Cov\left(\hat{\beta}_j, \hat{\beta}_k\right) = \left(X_j' X_j\right)^{-1} X_j' Cov\left(\varepsilon_j, \varepsilon_k\right) X_k \left(X_k' X_k\right)^{-1}$$

where the betas and epsilons are from least-squares regressions of Y_j on X_j and Y_k on X_k. One could use the simple estimator

$$Cov\left(\varepsilon_j, \varepsilon_k\right) = \lim_{n\to\infty} \frac{1}{n} \sum_i \hat{\varepsilon}_{ji} \hat{\varepsilon}_{ki}$$

but we instead use the heteroskedasticity-consistent Eicher-White formulation,

$$X_j' Cov\left(\varepsilon_j, \varepsilon_k\right) X_k = \lim_{n\to\infty} \frac{1}{n} \sum_i \hat{\varepsilon}_{ji} \hat{\varepsilon}_{ki} X_{ji} X_{ki}'.$$

Alternatively, the covariance matrix among the nine weeks' DID estimates could also be obtained using the nonparametric bootstrap.

[19] The first of three weeks prior to the start of campaign #2 overlaps with the week following campaign #1 for three days (see footnote 15) and that one of campaign #2's second "week" is actually only three days, since the campaign was only ten days long (see footnote 16).

these advertisements to the advertiser was R$33,000. Thus, our point estimate says that the total revenue benefit of the ads was nearly eight times the cost of the campaign, while even the lower bound of our 95% confidence interval indicates a revenue benefit of two times the cost (where a retailer with a 50% margin would approximately break even). Even more strongly than before, we conclude that "retail advertising works" in this case study.

We perform a specification test for each of the weekly DID estimates, similar to the specification test computed above for the 2-week DID estimate. This test determines whether the control group and the untreated members of the treatment group might pursue different time-varying purchasing behavior, which would invalidate our DID estimator's strategy of pooling these two groups. We present the results of the weekly estimates of this difference in Figure 9. During each of the 9 weeks following the start of campaign #1, the difference in time-series differences between control and untreated treatment group members fails to reject the null hypothesis that the DID model is correctly specified.

C. Summary of Persistence Results

To summarize the main result of this section, we find that the retail image advertising in this experiment led to persistent positive effects on sales for a number of weeks after the ads stopped showing. When we take these effects into account, we find a large return to advertising for the period of our sample. It is possible that we are still underestimating the returns to advertising, because our sales data end one week after the end of campaign #2 and, hence, our estimates omit any persistent effects that the advertising may have beyond the end of our sample period. We hope to further investigate display advertising's persistence in future experiments with longer panels of sales data.

VI. Detailed Results for Campaign #1

In this section, we dig deeper into several other dimensions of the data for the first campaign. For this purpose, we restrict attention to the three weeks of sales analyzed in Section V.A above. The questions we address in the present section are most intuitively asked about a single campaign, and we choose here to focus on the first, larger, and more impactful of the two campaigns in our experiment. Note that the second campaign cannot be analyzed cleanly on its

own, because the treatment and control groups were not re-randomized between campaigns, and as shown in the previous section, effects may be persistent across time. [20]

Despite the evidence on persistent effects in the previous section, we also know that longer time differences produce more scope for error in a difference-in-differences specification. Therefore, to be conservative, we ignore possible sales impacts more than a week after the campaign. In this section, we employ a difference in differences for three weeks before and after the start of campaign #1, comparing treated versus untreated individuals.

First, we decompose the effects of online advertising into offline versus online sales, showing that more than 90% of the impact is offline. We also demonstrate that most of the substantial impact on in-store sales occurs for users who merely view the ads but never click them. Second, we examine how the treatment effect varies with the number of ads viewed by each user. Third, we decompose the effects of online advertising into the probability of a transaction versus the size of the purchase conditional on a transaction.

A. Offline versus Online Sales and Views versus Clicks

In Table 7, we present a decomposition of the treatment effect into offline and online components by running the previous difference-in-differences analysis separately for offline and online sales. The first line of the table shows that the vast majority of the treatment effect comes from brick-and-mortar sales. The treatment effect of R$0.166 per treated individual turns out to consist of a R$0.155 effect on offline sales plus a R$0.011 effect on online sales. In other words, 93% of the treatment effect of the online ads occurred in *offline* sales. This result will be surprising to those who assume that online ads have an impact only on online sales.

In online advertising, the click-through rate (CTR) is a standard measure of performance. This measure (approximately 0.3% for the ads in this experiment) provides more information than is available in traditional media, but it still does not measure the variable that advertisers actually care most about: the impact on sales. An interesting question is, therefore, "To what extent do clicks on ads predict increases in retail sales due to advertising?"

[20] Our experiment was initially designed to examine the two campaigns individually, assuming that the effects would not persist for more than one week after each campaign. We appended the two datasets together only after it became clear that longer persistence was possible. This produced the overlap in days discussed in footnote 15. Restricting attention in this section to just three weeks after the first campaign allows us to avoid both expositional and statistical complications.

We answer this question in the second and third lines in Table 7. We partition the set of treated individuals into those who clicked on an ad (line 2) versus those who merely viewed ads but did not click any of them (line 3). Of the 814,000 individuals treated with ads, 7.2% clicked on at least one ad, while 92.8% merely viewed them. With respect to total sales, we see a treatment effect of R$0.139 on those who merely view ads, and a treatment effect of R$0.508 on those who click them. Our original estimate of the treatment effect can be decomposed into the separate effects for viewers versus clickers, using their relative weights in the population: R$0.166 = (92.8%)(R$0.139) + (7.2%)(R$0.508). The first component—the effect on those who merely view but do not click ads—represents 78% of the total treatment effect. Thus clicks, though the standard performance measure in online advertising, fail to capture the vast majority of the effects on sales.

The click-versus-view results are qualitatively different for offline than for online sales. For offline sales, those individuals who view but do not click ads purchase R$0.150 more than untreated individuals (a statistically significant difference). For online sales, the effect of viewing but not clicking is precisely measured to be near zero, so we can conclude that those who do not click do not buy online. In contrast, those who click show a large difference in purchase amounts relative to untreated individuals in both offline and online sales: R$0.215 and R$0.292, respectively. While this treatment effect for clickers is highly statistically significant for online sales, it is insignificant for offline sales due to a large standard error.

B. How the Treatment Effect Varies with the Number of Ads

We saw in Figure 2 that different individuals viewed very different numbers of ads during campaign #1. We now ask how the treatment effect varies with the number of ads viewed.

We wish to produce a smooth curve showing how this difference varies with the number of ads. Recall that for each individual, we observe the pre-post difference in purchase amounts (three weeks before versus three weeks after the start of the campaign). We perform a nonparametric, locally linear regression on this difference, using an Epanechnikov kernel with a bandwidth of 15 ad views. For readability, because the pre-post differences are negative on average and because we expect the treatment effect to be zero for those who did not view ads, we normalize the vertical intercept of the graph so that it equals zero for those with zero ad views.

Figure 10 gives the results, together with 95% confidence-interval bands around the conditional mean. We see that the treatment effect is initially increasing in the number of ads viewed. The effect peaks at approximately 50 ads viewed, for a maximum treatment effect of R$0.25, and remains almost flat at this level until it reaches 100 ad impressions per person. Beyond this point, the data becomes so sparse (only 6.1% of the treatment group sees more than 100 ad views) that the effect is no longer statistically distinguishable from zero.

We caution that one should not assume that this graph shows the causal effects on sales of increasing the number of ads shown to a given individual. This causal interpretation could be invalid because the number of ad views does not vary exogenously by individual. Each individual has a browsing "type" that determines the distribution of pages they will visit on Yahoo!, and this determines the average number of ads that user will receive. We know from the previous results in Table 4 that browsing behavior is correlated with the retail purchases in the absence of advertising on Yahoo!, so we shy away from the strongly causal interpretation we might like to make. We are on solid ground only when we interpret the graph as displaying heterogenous treatment effects by the user's browsing type.

The upward-sloping line on the graph represents our estimate of the cost to the retailer of purchasing a given number of ad impressions per person. This line has a slope of approximately R$0.001, which is the price per ad impression to the retailer. Thus, the graph plots the nonlinear revenue curve versus the linear cost curve for a given number of advertisements delivered to a given individual. The crossover that occurs at approximately 100 ad views is a breakeven point for revenue. For those individuals who viewed fewer than 100 ads (93.9% of the treatment group), the increased sales exceed the cost of the advertisements.

If we want to look at incremental profits rather than incremental revenues, we could assume a 50% retail profit margin and multiply the entire benefit curve by 50%, effectively reducing its vertical height by half. Given the shape of the curve, the breakeven point remains approximately the same, at around 100 ads per person. For the 6% of individuals who received more than 100 ads, the campaign might not have been cost-effective, though statistical uncertainty prevents this from being a firm conclusion. The retailer might be able to gain from a policy that caps the number of ad views per person at 100, because it avoids spending money on individuals for whom there does not appear to be much positive benefit. This hypothesis could fruitfully be investigated in future experiments.

C. Probability of Purchase versus Basket Size

So far, our analysis has focused on advertising's effects on the average purchase amount per person, including those who made purchases as well as those who did not. We can decompose the effects of advertising into two separate channels of interest to retailers: the effect on the probability of a transaction, and the effect on "basket size," or purchase amount conditional on a transaction. To provide some base numbers for reference, during the three-week period after the start of the campaign, individuals treated with ads had a 6.48% probability of a transaction, and the average basket size was R$40.72 for those who purchased. The product of these two numbers gives the average (unconditional) purchase amount of R$2.64 per person. We reproduce these numbers in the last column of Table 8 for comparison to our treatment-effect results.

In Table 8, the first column shows the estimates of the treatment effects on each variable of interest. As before, our treatment effects come from a difference in differences, comparing those treated with ads versus those untreated, using three weeks of data before and after the start of the campaign.

First we investigate advertising's impact on the probability of a transaction,[21] with results shown in the first row of the table. We find an increase of 0.102% in the probability of purchase as a result of the advertising, and the effect is statistically significant (p=0.03). This represents an increase of approximately 1.6% relative to the average probability of a transaction.

Next, we consider the effect on basket size. In this application, we wish to analyze the magnitude of a purchase conditional on a transaction. Since sales data are sparse and most purchasers do not purchase in both time periods, we cannot employ the same DID estimator as before, running a regression with the dependent variable of time differences in sales by individual. Instead, we compute DID using group means of basket size, and to compute a consistent standard error we pay careful attention to possible time-series correlation.[22] As shown in the second row of Table 8, the advertising campaign produced an increase in basket size of

[21] We include negative purchase amounts (net returns) as transactions in this analysis. Since we previously found that advertising decreases the probability of a negative purchase amount, the effect measured here would likely be larger if we restricted our analysis to positive purchase amounts.

[22] When comparing the mean time-series difference for treated individuals to the mean time-series difference for untreated individuals, we know those two means are independent, so standard errors are straightforward. But when computing a difference in differences for four group means, we know we should expect correlation between pre-campaign and post-campaign basket size estimates since some individuals purchase in both periods and may have serially-correlated sales.

R$1.75, which is statistically significant (p=0.018). Compared with the baseline basket size of $40.72, this represents an increase of approximately 4.5%.

To summarize, we initially found that the treatment caused an increase of R$0.166 in the average (unconditional) purchase amount. This decomposes into an increase of 0.102% in the probability of a transaction, as well as an increase of R$1.75 in the purchase amount conditional on a transaction, representing percentage increases relative to baseline of 1.6% and 4.5% respectively. Thus, we estimate that about one-fourth of the treatment effect appears to be due to increases in the probability of a transaction, and about three-fourths due to increases in basket size.

VII. Conclusion

Despite the economic importance of the advertising industry, the causal effects of advertising on sales have been extremely difficult to quantify. In this study, we take a substantial step forward in this measurement problem by conducting a large-scale field experiment that systematically varies advertising to a subject pool of over one million retail customers on Yahoo! With such a large individual-level dataset, we are just on the frontier of being able to measure economically meaningful effects: our power calculations show that for a standard 5% two-sided hypothesis test, even an advertising campaign that doubles the advertiser's money in the short run would only be detected with probability 63%. Sales at this retailer have high variance and this online advertising campaign is just one of many factors that influence purchases, which makes treatment-control differences rather noisy. For more precise estimates, we employ a difference-in-difference estimator using panel data on weekly individual transactions, exploiting both experimental and non-experimental variation in advertising exposure. This DID estimator requires more assumptions than the simple treatment-control difference estimator, but the two estimators provide similar results and the DID estimator passes a specification test on the assumptions.

Our primary result is that in this case study, retail advertising works! We find positive, sizeable, and persistent effects of online retail advertising on retail sales. The ad effects appear to persist for several weeks after the last ad has been shown. In total, we estimate that the retailer gained incremental revenues more than seven times as large as the amount it spent on the online ads in this experiment.

Though some people assume that online advertising has most of its effects on online retail sales, we find the reverse to be true. This particular retailer records 86% of its sales volume offline, and we estimate 93% of our treatment effect to occur in offline sales. Online advertising has a large effect on offline sales.

Furthermore, though clicks are a standard measure of performance in online-advertising campaigns, we find that online advertising has even more substantial effects on the set of people who merely view the ads than on the set who actually click them. Clicks are a good predictor of online sales, but not of offline sales. We decompose the total treatment effect to show that 78% of the lift in sales comes from those who view ads but do not click them, while only 22% can be attributed to those who click.

We find that the treatment effect of advertising is largest for those individuals who browsed Yahoo! enough to see between 25 and 100 ad impressions during a two-week period. We also find that online advertising increases both the probability of purchase and the average purchase amount, with about three-quarters of the treatment effect coming through increases in the average purchase amount.

Another important result is a demonstration of how poorly one can measure the causal effects of advertising using common modeling strategies. If we had neither an experiment nor panel data available to us, but instead attempted to estimate these effects using cross-sectional variation in endogenous advertising exposure, we would have obtained a result that was opposite in sign to the true estimate. The magnitude of the selection bias would be more than three times the magnitude of the true measured effect of advertising. We also show that being more careful with the non-experimental data can produce more accurate results, as implementing a difference in differences on the panel data gives us a very similar estimate to the experimental measurement. In fact, by supplementing our experimental variation with the non-experimental before-after variation in sales, we obtain a more efficient estimate of the treatment effect (partially making up for our inability to observe the part of the control group who would have been treated).

In future research, we hope to replicate these results with other retailers. We are using what we have learned in this study in order to design better experiments: for example, future experiments will carefully mark control-group members who would not have browsed in ways that exposed them to ads, so that we can more efficiently estimate the treatment effect on the

treated in the experiment. We also wish to investigate related factors in online advertising, such as the value of targeting customers with particular demographic or online-browsing-behavior attributes that an advertiser may think desirable. The ability to conduct a randomized experiment with a million customers and to match individual-level sales and advertising data makes possible exciting new measurements about the economic effects of advertising, and we look forward to additional explorations on this new frontier.

References

Abraham, M. 2008. "The Off-Line Impact of Online Ads." *Harvard Business Review*, 86(April): 28.

Abraham, M. and L. M. Lodish. 1990. "Getting the Most out of Advertising and Promotion." *Harvard Business Review*, 68(3): 50-60.

Ackerberg, Daniel. 2001. "Empirically Distinguishing Informative and Prestige Effects of Advertising." *RAND Journal of Economics*, 32(2): 316-333.

Ackerberg, Daniel. 2003. "Advertising, Learning, and Consumer Choice in Experience-Good Markets: An Empirical Examination." *International Economic Review*, 44(3): 1007-1040.

Anderson, Eric, and Duncan Simester. 2008. "Dynamics of Retail Advertising: Evidence from a Field Experiment." Forthcoming, *Economic Inquiry*.

Bagwell, K. 2008. "The Economic Analysis of Advertising." *Handbook of Industrial Organization*, vol. 3, Mark Armstrong and Robert Porter, eds. Amsterdam: Elsevier B.V., 1701-1844.

Berndt, Ernst R. 1991. *The Practice of Econometrics: Classic and Contemporary*. Reading, Massaschusetts: Addison-Wesley.

Cameron, A. C. and P. K. Trivedi. 2005. *Microeconometrics*. Cambridge University Press.

Dorfman, R. and P. O. Steiner. 1954. "Optimal Advertising and Optimal Quantity," *American Economic Review*, 44(5): 826-836.

Hu, Y., L. M. Lodish, and A. M. Krieger. 2007. "An Analysis of Real World TV Advertising Tests: a 15-Year Update." *Journal of Advertising Research*, 47(3): 341-353.

Lewis, Randall A. 2010. "Where's the 'Wear-Out?' Online Display Ads and the Impact of Frequency," Yahoo! Research working paper.

Lewis, Randall A. and David H. Reiley. 2010. "Advertising Especially Influences Older Users: A Yahoo! Experiment Measuring Retail Sales," Yahoo! Research working paper.

Levitt, Steven, and John A. List. 2009. "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review*, 53(1): 1-18.

Lodish, L. M., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., and M. E. Stevens. 1995a. "How T.V. Advertising Works: a Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments." *Journal of Marketing Research*, 32(2): 125-139.

Lodish, L. M., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., and M. E. Stevens. 1995b. "A Summary of Fifty-Five In-Market Experiments of the Long-Term Effect of TV Advertising." *Marketing Science*, 14(3): 133-140.

Schmalensee, Richard. 1972. *The Economics of Advertising*. Amsterdam: North-Holland.

Tables and Figures

Table 1 - Summary Statistics for the Campaigns

	Campaign 1	Campaign 2	Both Campaigns
Time Period Covered	Early Fall '07	Late Fall '07	
Length of Campaign	14 days	10 days	
Number of Ads Displayed	32,272,816	9,664,332	41,937,148
Number of Users Shown Ads	814,052	721,378	867,839
% Treatment Group Viewing Ads	63.7%	56.5%	67.9%
Mean Ad Views per Viewer	39.6	13.4	48.3

Figure 1– Yahoo! Front Page with Large Rectangular Advertisement

Table 2 - Basic Summary Statistics for Campaign #1

	Control	Treatment
% Female	59.5%	59.7%
% Retailer Ad Views > 0	0.0%	63.7%
% Yahoo Page Views > 0	76.4%	76.4%
Mean Y! Page Views per Person	358	363
Mean Ad Views per Person	0	25
Mean Ad Clicks per Person	0	0.056
% Ad Impressions Clicked (CTR)	-	0.28%
% Viewers Clicking at Least Once	-	7.2%

Figure 2 - Ad Views Histogram

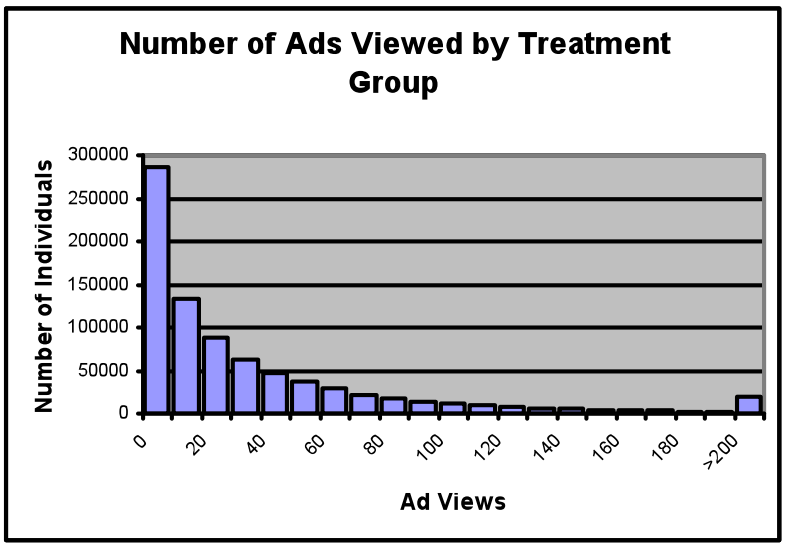

Table 3 - Weekly Sales Summary

		Mean Sales	Std. Dev.	Min	Max	Transactions
Campaign #1						
09/24	3 Weeks Before	**R$ 0.939**	14.1	-932.04	4156.01	42,809
10/01	2 Weeks Before	**R$ 0.937**	14.1	-1380.97	3732.03	41,635
10/08	1 Week Before	**R$ 0.999**	14.3	-1332.04	3379.61	43,769
10/15	Week 1 During	**R$ 0.987**	13.5	-2330.10	2163.11	43,956
10/22	Week 2 During	**R$ 0.898**	13.3	-1520.39	2796.12	40,971
10/29	Week 1 Following	**R$ 0.861**	13.3	-1097.96	3516.51	40,152
Campaign #2						
11/02	3 Weeks Before	**R$ 1.386**	16.4	-1574.95	3217.30	52,776
11/09	2 Weeks Before	**R$ 1.327**	16.6	-654.70	5433.00	57,192
11/16	1 Week Before	**R$ 0.956**	13.4	-2349.61	2506.57	45,359
11/23	Week 1 During	**R$ 1.299**	16.7	-1077.83	3671.75	53,428
11/30	Week 2 During (3 Days)	**R$ 0.784**	14.0	-849.51	3669.13	29,927
12/03	Week 1 Following	**R$ 1.317**	16.1	-2670.87	5273.86	57,522

N=1,577,256 observations per week

Figure 3 - Offline and Online Weekly Sales

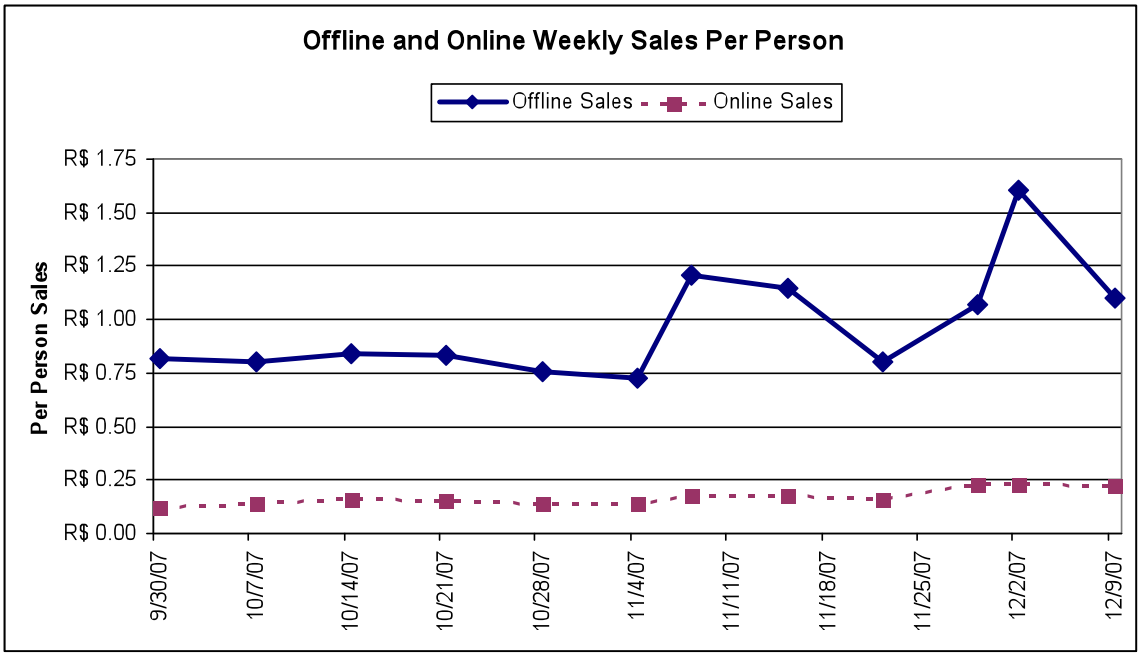

Table 4 - Two Week Treatment Effect Offline/Online Decomposition

	Before Campaign (2 weeks) Mean Sales/Person	During Campaign (2 weeks) Mean Sales/Person	Difference (During – Before) Mean Sales/Person
Control:	R$ 1.95	R$ 1.84	-R$ 0.10
	(0.04)	(0.03)	(0.05)
Treatment:	1.93	1.89	-R$ 0.04
	(0.02)	(0.02)	(0.03)
Exposed to Retailer's Ads:	1.81	1.81	R$ 0.00
	(0.02)	(0.02)	(0.03)
Not Exposed to Retailer's Ads:	2.15	2.04	-R$ 0.10
	(0.03)	(0.03)	(0.04)

Figure 4 - Histogram of Campaign #1 Sales by Treatment and Control

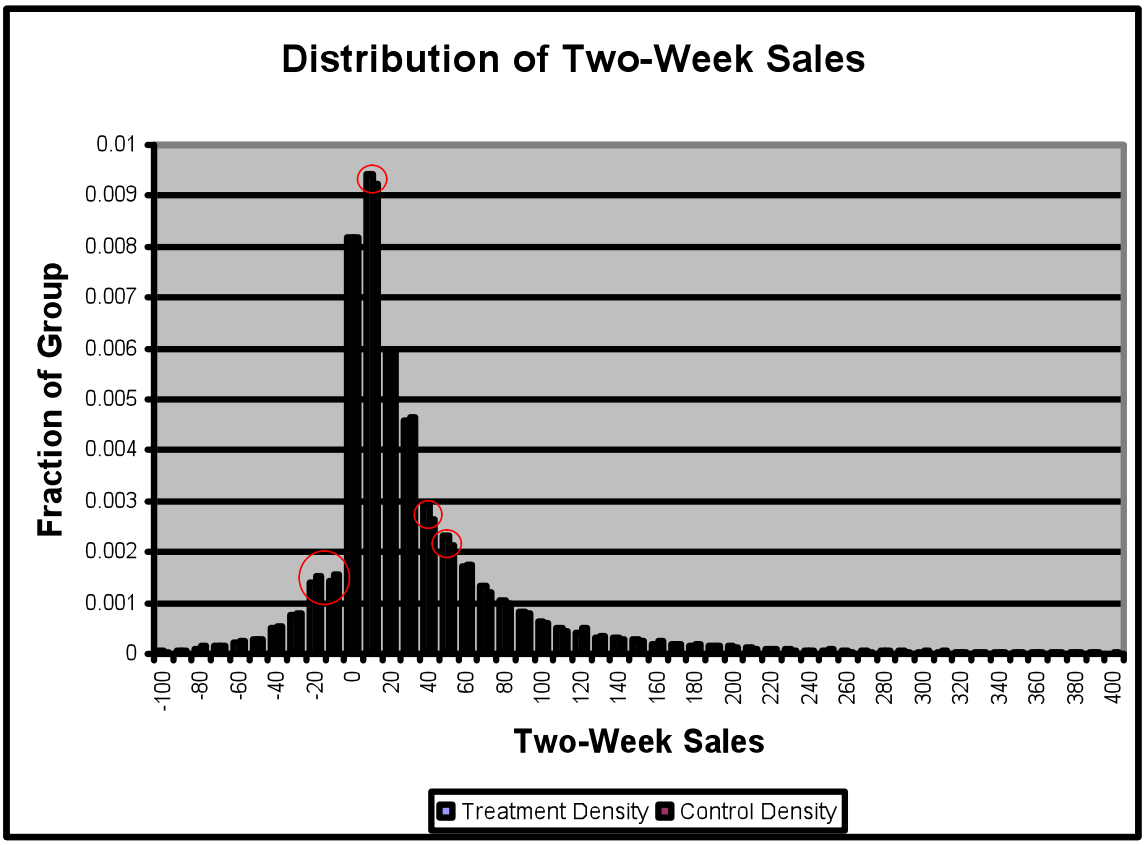

Figure 5 - Difference between Treatment and Control Sales Histograms

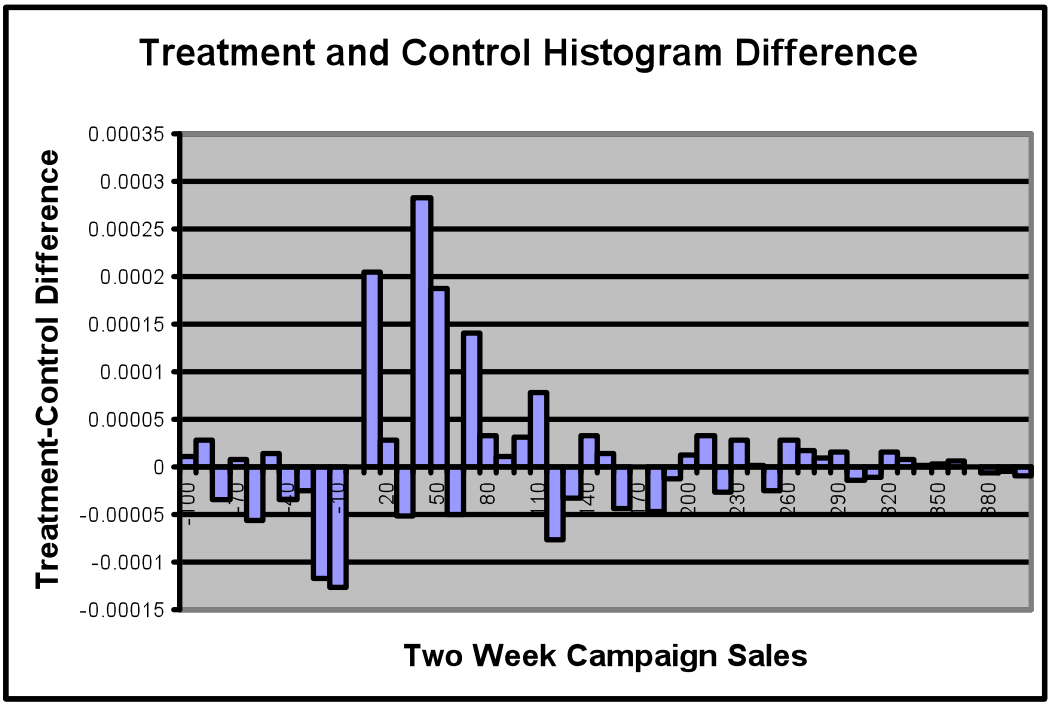

Figure 6 - Histogram of Difference in Three-Week Sales for Treated and Untreated Groups

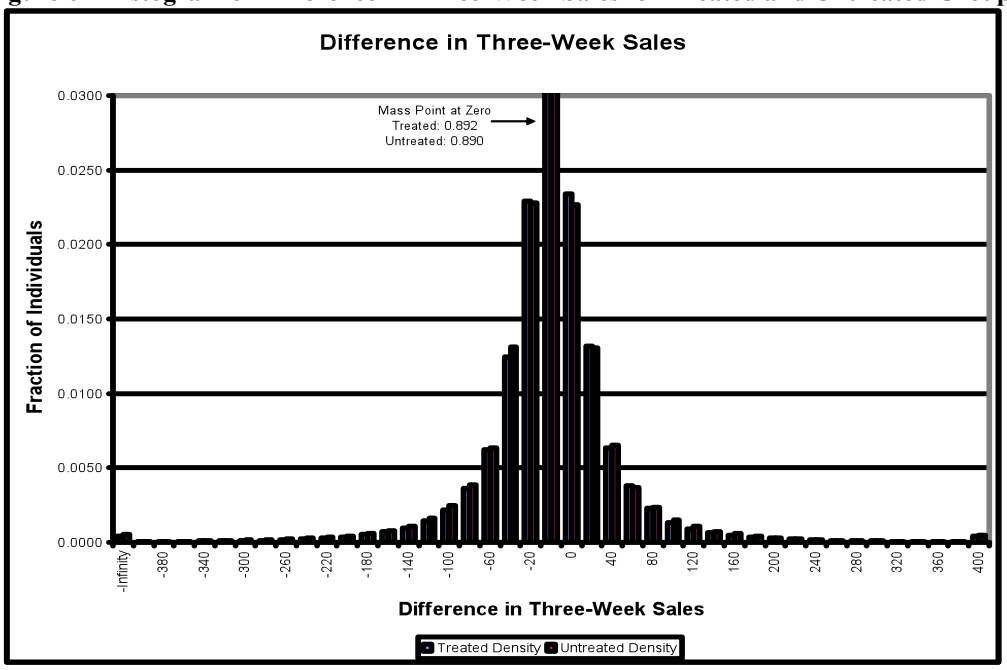

Figure 7 - Difference in Treated and Untreated Three-Week Sales Histograms

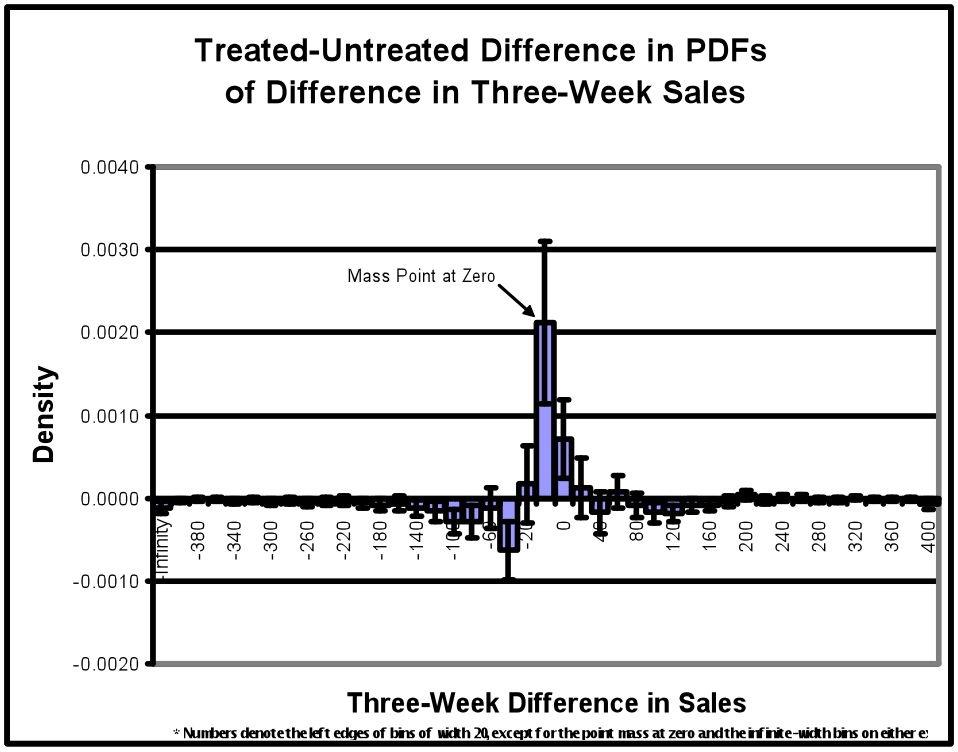

Table 5 - Weekly Summary of Effect on the Treated

	Treatment Effect*	Robust S.E.
Campaign #1		
Week 1 During	**R$ 0.047**	0.024
Week 2 During	**R$ 0.053**	0.024
Week 1 Following	**R$ 0.061**	0.024
Campaign #2		
3 Weeks Before	**R$ 0.011**	0.028
2 Weeks Before	**R$ 0.030**	0.029
1 Week Before	**R$ 0.033**	0.024
Week 1 During	**R$ 0.052**	0.029
Week 2 During (3 Days)	**R$ 0.012**	0.023
Week 1 Following	**R$ 0.004**	0.028
	N=1,577,256 obs. per week	

* For purposes of computing the treatment effect on the treated, we define "treated" individuals as having seen at least one ad in either campaign prior to or during that week.

Figure 8 - Weekly DID Estimates of the Treatment Effect for Both Campaigns

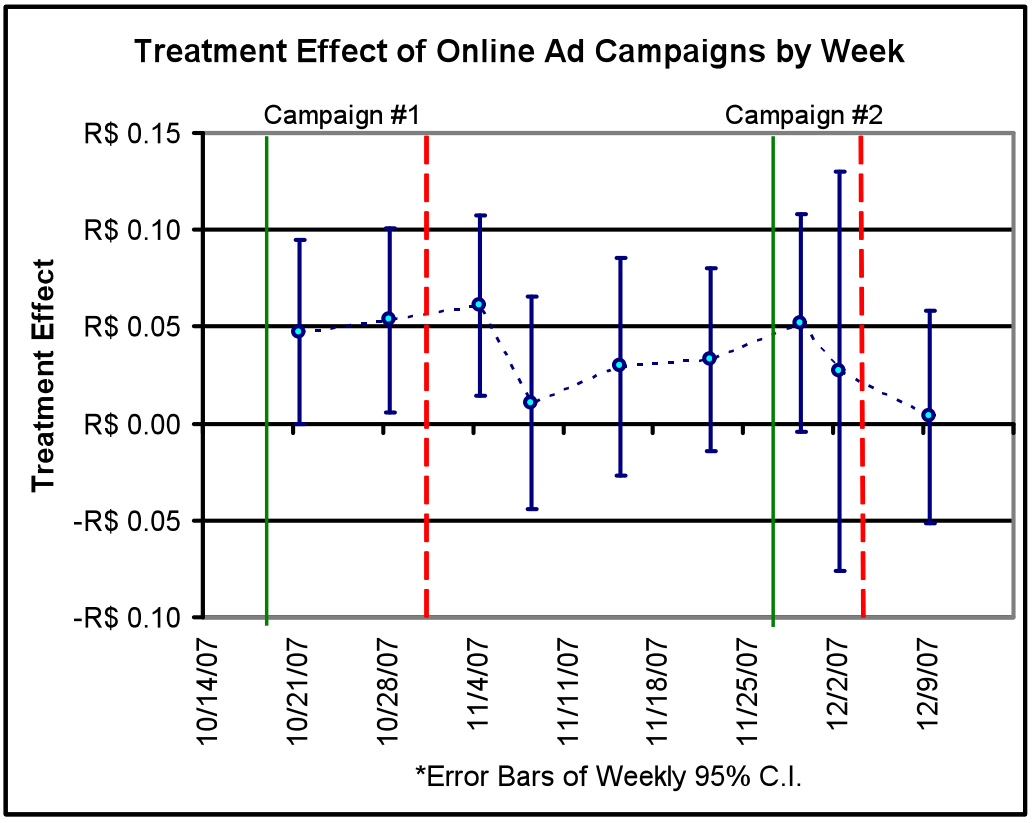

Table 6 - Results Summary for Both Campaigns

	Treatment Effect	Robust S.E.	t-stat	P(t>T)
Average Weekly Effect				
Simple Average (OLS)	**R$ 0.035**	0.0155	2.28	0.011
Efficient Average (GLS)	**R$ 0.039**	0.0147	2.65	0.004
Cumulative Effects over Both Campaigns				
Cumulative Sales	**R$ 0.299**	0.123	2.42	0.008
Simple Aggregate Effect (OLS)	**R$ 0.282**	0.124	2.28	0.011
Efficient Aggregate Effect (GLS)	**R$ 0.311**	0.117	2.65	0.004
Length of Measured Cumulative Effects		8 weeks		

Figure 9 - Weekly DID Specification Test

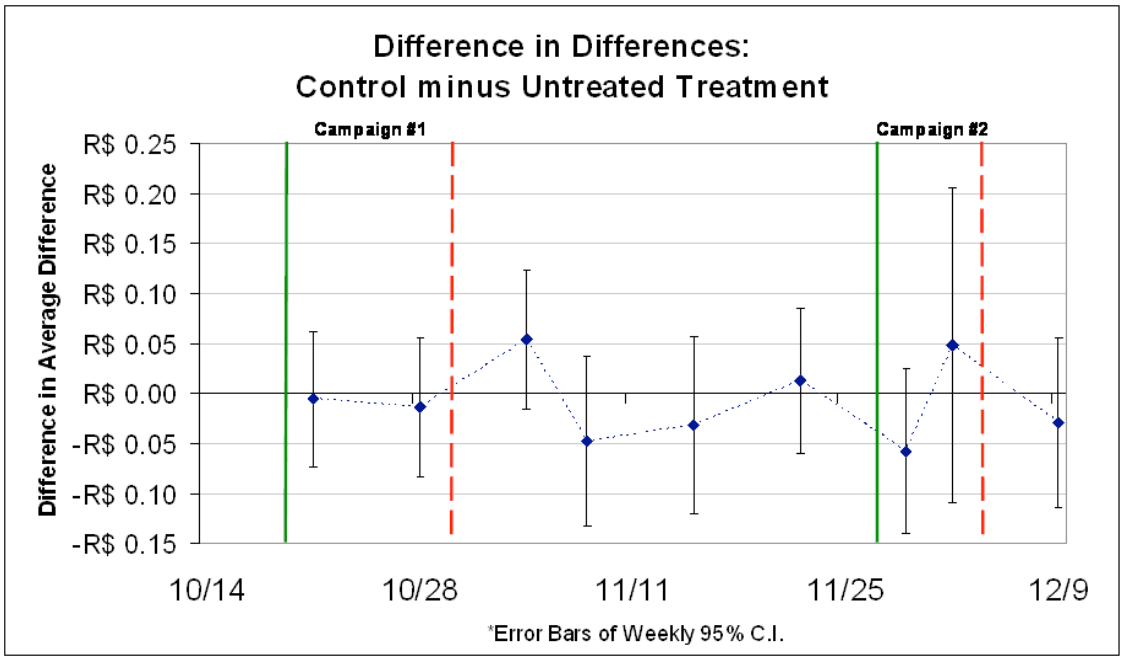

Table 7 - Offline/Online Treatment Effect Decomposition

	Total Sales	Offline Sales	Online Sales
Ads Viewed	**R$ 0.166**	**R$ 0.155**	R$ 0.011
[63.7% of Treatment Group]	(0.052)	(0.049)	(0.016)
Ads Viewed Not Clicked	**R$ 0.139**	**R$ 0.150**	-R$ 0.010
[92.8% of Viewers]	(0.053)	(0.050)	(0.016)
Ads Clicked	**R$ 0.508**	R$ 0.215	**R$ 0.292**
[7.2% of Viewers]	(0.164)	(0.157)	(0.044)

Figure 10 - Nonparametric Estimate of the Treatment Effect by Ad Viewing Outcome

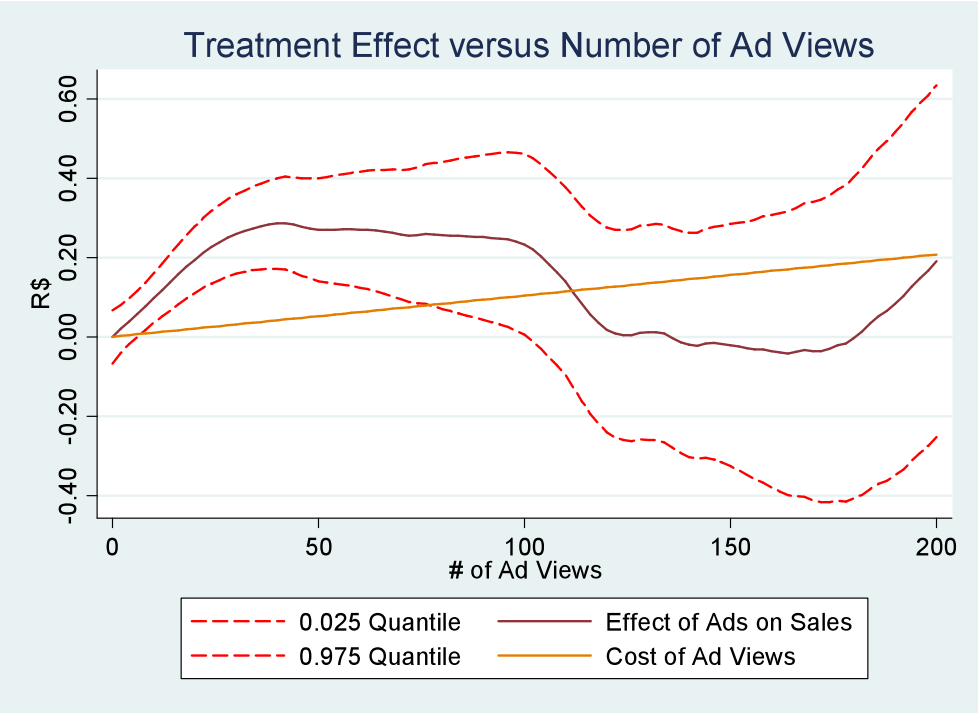

Table 8 - Decomposition of Treatment Effect into Basket Size and Frequency Effects

	3-Week DID Treatment Effect	Treated Group Level*
Pr(Transaction)	**0.102%** (0.047%)	**6.48%**
Mean Basket Size	**R$ 1.75** (0.74)	**R$ 40.72**
Revenue Per Person	**R$ 0.166** (0.052)	**R$ 2.639**

* Levels computed for those treated with ads during Campaign #1, using three weeks of data following the start of the campaign.