Retail Advertising Works!

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

Randall Lewis and David Reiley*

This Version: 7 January 2009

Abstract

A randomized experiment performed in cooperation between Yahoo! and a major retailer allows us to measure the effects of online advertising on sales. We exploit a match of over one million customers between the databases of Yahoo! and the retailer, assigning them to treatment and control groups for an online advertising campaign for this retailer and then measuring each individual's weekly sales at this retailer, both online and in stores. By combining a controlled experiment with panel data on purchases, 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 we estimate the total effect on revenues to be more than eleven times the retailer's expenditure on advertising during the study. Additional results explore differences in the number of advertising impressions delivered to each individual, age and gender demographics, online and offline sales, and the effects of advertising on those who click the ads versus those who merely view them. Our results provide the best measurements to date of the effectiveness of image advertising on sales, and we shed light on important questions about online advertising in particular.

---

* Lewis: Massachusetts Institute of Technology, randallL@mit.edu. Reiley: Yahoo! Research and University of Arizona, reiley@eller.arizona.edu. 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, Kei Hirano, John List, Preston McAfee, Paul Ruud, Michael Schwarz, Pai-Ling Yin, and seminar participants at University of Arizona, University of California at Davis, New York University, Sonoma State University, Vassar College, the FTC Microeconomics Conference, and Economic Science Association meetings in Pasadena, Lyon, and Tucson.

I. Introduction

The retailing pioneer John Wanamaker (1838-1922) famously remarked, "Half the money I spend on advertising is wasted; the trouble is I don't know which half." Measuring the impact of advertising on sales has remained a difficult problem for more than a century. A particular problem 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 a year 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.

In general, advertisers do not systematically vary their levels of advertising to measure the effects on sales.[1] 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 advertisers 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 arranged to 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 3000 households. Abraham and Lodish 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, 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

[1] Notable exceptions include direct-mail advertising, and more recently, search-engine 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.

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.

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).

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 three advertising campaigns in this experiment. In Section V, we demonstrate and measure the persistence of this effect after the campaign has ended. In Section VI, we 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 of ads, the effect by age and gender demographics, 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 whether individuals were exposed to a nationwide retailer's display-advertising campaign on Yahoo! We then measure the impact of the advertising on individuals' weekly purchases, both online and in stores. To achieve this end, we made use of matches between both postal and email addresses in the retailer's customer database and the addresses in Yahoo!'s user database. This match yielded a sample of 1,577,256 individuals.[2]

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

Of these matched users, we assigned 81% to a treatment group who subsequently viewed three 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 three different advertising campaigns over the course of four months, separated by approximately one-month intervals. Table 1 gives summary statistics for the three campaigns, which delivered 32 million, 10 million, and 17 million impressions, respectively. Figure 1 shows that by the end of the third campaign, a total of 924,000 users had been exposed to ads. These individuals viewed an average of 64 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. Yahoo! advertising turns out to explain a very small fraction of the variance in weekly sales, but because of the randomization, they are uncorrelated with any other influences on shopping behavior.

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, groups.yahoo.com, and maps.yahoo.com. Figure 2 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. All financial figures involving treatment effects and sales will be reported in R$, or "Retail Dollars," rather than US dollars.

III. Sales and Advertising Data

Table 2 provides some summary statistics that indicate a successful randomization. The treatment group was 59.7% female while the control group was 59.5% female, a statistically insignificant difference (p=0.212). During campaign #1, the proportion of individuals who did any browsing on the Yahoo! network was 76.4% in each group (p=0.537). The mean number of Yahoo! page views was 363 pages for the treatment group versus 358 in the control group, another statistically insignificant difference (p=0.627).

[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 treatment of viewing advertisements was delivered randomly by the Yahoo! ad server such that even though 76.4% of the treatment group visited a Yahoo! website, only 63.7% of the treatment group was actually shown the retailer's ads. On average, a visitor was shown the ads on only 7.0% of the pages they visited during Campaign #1, but the probability of being shown on any particular page depended on several factors including, but not limited to, the property they visited, specific content on the page, and user demographics.

The number of the ads viewed by each Yahoo! user in this campaign is quite skewed. The match between the Yahoo! data and the retailer's sales data should tend to reduce the number of non-human "bots" or automated browsing programs since a would-be "bot" would have to make a purchase at the retailer in order to be included in our sample. However, there still remains a very small percentage of users who have extreme browsing behavior. Figure 3 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 of the ads during the two weeks of the online ad campaign. The maximum number of the ads viewed during the campaign period by one individual was 6050.[4]

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 the proportion of the designated treatment group who clicked at least one ad in this campaign was 4.6%. Of those who actually saw at least one ad, the fraction who clicked at least one ad was 7.2%.

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 the week following each of the three campaigns. Sales amounts include all purchases that the retailer could link to each individual customer in the database, primarily by use of credit-card information.[5]

Table 3 provides a weekly summary of the sales data, while Figure 4 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 aggregated 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

[4] Although the data suggests extreme numbers of ads, Yahoo! engages in extensive "anti-click-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.

[5] To the extent that these customers make purchases (such as with cash) that cannot be tracked by the retailer, our estimate may underestimate the total effect of advertising on sales. However, the retailer believes that they track at least 90% of purchases for these customers.

than 5% of individuals in a given week make any transaction (see last column of Table 4). For those who do make a purchase, the transaction amounts exhibit large positive and negative amounts (the latter representing returns), but well over 90% of purchase amounts lie between –R$100 and +R$200.

We next look at the results of the experiment for the first of the three advertising campaigns. Throughout the paper we primarily focus on campaign #1 for several reasons. The first is that more than 60% of the ads during all three campaigns were shown during those two weeks. Second, subsequent campaigns were shown to the same treatment and control groups, preventing us from examining any before and after differences. Third, even with 1.6 million customers, the expected magnitudes of any treatment effect that depends on frequency are likely too small for us to have the statistical power to estimate given the high volatility of the data. With these issues taken into consideration, we examine the effects of campaign #1.

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. However, with online display advertising, it is very easy to track potential customers and bill an advertiser by the number of impressions an ad is delivered. Although there is a significant difference between a delivered ad and a seen ad, the ability to count the number of attempted exposures to an individual allows us to investigate the treatment effect in many ways such as by focusing on those who were delivered at least one ad impression.

Table 5 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. 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 suggests a positive treatment effects of ads of approximately R$0.05 per person, but the effect is not statistically significant at conventional levels (p=0.162).

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. This technique gives a slightly larger estimate of R$0.06 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. (Recall that they did not see ads because of their

individual browsing choices.) Ideally, we would remove these 36% of individuals both from the treatment and control groups in order to get an estimate of the advertising treatment on those who could be treated. Unfortunately, we are unable to determine which control-group members would have seen ads for this campaign had they been in the treatment group.[6] Instead of removing these individuals, we scale up our diluted treatment effect (R$0.05) by dividing by 0.64, the fraction of individuals treated.[7] This gives us an estimate of the treatment effect on those treated with ads: R$0.083. The standard error is also scaled proportionally, leaving the level of statistical significance unaffected (p=0.162).

The last two rows of the table show us an interesting difference between those treatment-group members who saw ads in this campaign and those who didn't. Before the campaign, 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) is evidence of heterogeneity in shopping behavior that happens to be correlated with ad views (through Yahoo! browsing behavior). That is, for this population of users, those who browse Yahoo! more actively also have a tendency to purchase less at the retailer, independent of the number of ads shown. Therefore, 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). The pre-campaign data are crucial in allowing us to demonstrate the magnitude of this possible selection bias.

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. (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

---

[6] 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, and we were unfortunately unable to observe exactly which control-group members might have seen ads for this campaign.

[7] 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.

individuals' sales did. This will lead us to our preferred estimator below, a difference in differences between treated and untreated individuals (where "untreated" includes both control-group members and untreated members of the designated treatment group).

Next we look at the shape of the distribution of sales. Figure 5 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.[8] 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 6 plots the difference between the two histograms in Figure 5. The treatment effect is the average over this difference between treatment and control sales distributions.

Above, we noted that 36% of the treatment group did not see ads, and we are unable to identify the corresponding 36% of the control group who would not have seen ads, in order to remove both groups from the analysis. However, we were able to obtain the total number of pages browsed on the Yahoo! network during the campaign for each individual in both the treatment and control groups.[9] We know that someone who viewed no pages would not have viewed ads for this retailer, though the converse is not also true. We find that 76.4% of both treatment and control groups had nonzero page views on the Yahoo! network during the campaign.

Table 5 shows our results for those individuals observed to have page views during the campaign. Excluding the 23.6% of individuals who did not browse the Yahoo! network, we obtain a statistically significant treatment-control difference of R$0.078. Even this result is somewhat diluted by individuals who did not actually view ads. We know that 63.7%/76.4% = 83.4% of those who saw pages actually saw this retailer's ads, so we need to scale up the treatment effect again by dividing by 0.834. This yields an average effect on the treated of R$0.093, which is marginally statistically significant at the same level (p=0.09).

---

[8] We trim about 400 observations from the left and 400 observations from the right, of a total of 75,000 observations with positive 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.

[9] The original dataset did not contain data on individuals' page views, so including this variable required a data merge. Some observations were not uniquely matched using available matching variables. All page view values for these observations were attached to an observation that had matching values for the matching variables, but may not have been precisely the same observation. To handle data analysis correctly on this imperfectly merged data, we grouped together all imperfectly matched observations were (inefficiently) grouped together and all independent variables were averaged in order to eliminate this measurement error. We note that when two observations are mismatched, if we average their independent variables, we eliminate the measurement error of that mismatch since their measurement error is perfectly negatively correlated. By doing this, we preserve the unbiasedness of our least-squares regressions. For more details, see Lewis (2008).

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 equation:

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

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 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 beta coefficient is directly comparable to the previous "rescaled" estimates, as it measures 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<0.01$) 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. Since the cost of campaign #1 was approximately R$20,000[10], the point estimate suggests that the ads produced four times as much revenue as they cost the retailer. From this follows our conclusion that "retail advertising works!"

[10] 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!

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 uncorrelated with the DID error term. This assumption could be violated if some external event at some point during the experiment had different effects on the retail purchases of those who did and did not see ads. For example, perhaps in the middle of the time period studied, the retailer did a direct-mail campaign we don't 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 experimental estimates are very similar in magnitude to the DID estimates: R$0.083 for the simple comparison of levels between treatment and control, R$0.093 for the same estimate restricted to those individuals with positive Yahoo! page views, and R$0.102 for the DID estimate.

The similarity between these three different 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).

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, only 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 which, as previously mentioned, include 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 after the 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 7 of the distributions of three-week pre-post sales differences. Note three differences between Figure 7 and the histogram presented earlier in Figure 5: (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 1, 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 measure via difference in differences.

The three-week DID results can be found in the final column of Table 6, with the two-week results included for comparison. 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.

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 Campaigns #2 and #3. Recall that for each campaign, 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, for example, 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.[11]

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 figure will be the weekly average of sales in the three weeks preceding the start of Campaign #1. Our "post-period" sales figure will be the sales during a given week after the start of Campaign #1. We then

[11] 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 the campaign" and the "fourth week after the start of the campaign." 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.

compute a separate DID estimate for each week, beginning with the first week of Campaign #1 and ending with the week following Campaign #3.[12,13,14]

The results can be found in Table 7, represented graphically in Figure 9. In the figure, vertical lines indicate the beginning (solid) and end (dashed) of each of the three campaigns. The estimated treatment effects in later weeks thus include cumulative effects of all campaigns run to date. The average weekly treatment effect on the treated is R$0.045, with individual weekly estimates ranging from R$0.004 to R$0.080. Some of the individual weekly treatment effects are statistically indistinguishable from zero (95% confidence intervals graphed in Figure 9), but, strikingly, every single one of the point estimates is positive. We particularly note the large, positive effects estimated during the inter-campaign periods, as many as four weeks after ads stopped showing for this retailer on Yahoo!

To obtain an estimate of the cumulative effect of all three campaigns, we use all fifteen weeks of data. We present the results of two different methods in Table 8. The first method involves an estimate of the average weekly treatment effect, comparing the average of the fifteen weeks after Campaign #1 to the average of the three weeks prior to campaign #1. The estimate is R$0.045. We then multiply this estimate by the number of independent weeks of sales data in the sample period, which is actually thirteen weeks and three days. This estimate gives us an average treatment effect of the ads (on those who saw at least one ad during one of the three campaigns) of R$0.532 with a standard error of R$0.196.

A more econometrically efficient method is to compute an average of the 15 weekly estimates of the treatment effect, taking care to report standard errors that account for the covariances between regression coefficients across weeks. The table reports an optimally weighted average of the 15 per-week treatment effects,[15] with a simple average included

---

[12] There is an 11-day period of unobserved sales between Campaign #2 and Campaign #3.

[13] Because two of the campaigns lasted ten days rather than an even number of weeks, in two cases the second "week" of a campaign consists of only three days instead of seven. In the cases of 3-day "weeks," we scale up the sales data by 7/3 to keep consistent units of sales per week.

[14] Unfortunately, data for Campaign #3 came in a separate file from the third-party matching service, without unique identifiers, so we could not link it directly to the data for Campaigns #1 and #2. We imperfectly matched the Campaign #3 data back in, using a match on observables that were common across data sets. In the merge 80% of observations are uniquely matched and the remaining 20% are matched into groups. The combined number of groups and uniquely matched observations is 1,385,255. See footnote 9 and Lewis (2008) for details on the GLS regression we used to weight the group matches appropriately.

[15] 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 fifteen regression coefficients. These covariances can be computed 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 are from least squares regression coefficients from regressing Y_j on X_j and Y_k on X_k. We estimate $Cov(\varepsilon_1,\varepsilon_2)$ from the residuals of each regression. 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 Eicher-White formulation:

for comparison. The weighted average is R$0.048 (R$0.014). We then multiply this figure by 13.43 to get a total effect across the entire time period of observation, since the "fifteen-week" time period actually includes a total of only thirteen weeks and six days. This multiplication gives us a figure of R$0.645 (R$0.183). This is slightly larger than the total effect reported in the previous paragraph, because it assumes that all users were treated from the beginning of campaign #1.

To estimate the total benefit of the three campaigns, we take our estimate of R$0.645 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 864,000 (see the graph in Figure 1). This gives us a 95% confidence interval estimate of the total incremental revenues due to ads of R$560,000 ± 310,000. For comparison, the total cost of these advertisements to the advertiser was R$51,000. Thus, our point estimate says that the total revenue benefit of the ads was more than eleven times the cost of the campaign, and even the lower bound of our confidence interval indicates a benefit of more than six times the cost. Even more strongly than before, we conclude that retail advertising works!

Similar to the specification test computed for the DID estimate during the first 3 weeks, we consider a similar specification test to determine whether the control group and the untreated treatment group individuals pursue different time-varying purchasing behavior. We present the results of the weekly estimates of this difference in Figure 10. During the first 9 weeks, there is no sizeable aberration to suggest rejection of the null. In fact, the average difference is that the control group individuals on average purchase less. Since the control group is composed of individual who would and would not have seen ads, for the subgroup that saw ads, we would infer that their change in purchase behavior would have been even less than the untreated treatment group individuals. This suggests that DID may slightly underestimate the treatment effect during those weeks. However, during the last 6 weeks, there seems to be some evidence that the DID estimates actually overestimate the treatment effect. While this may be problematic for performing inference on the weekly treatment effects, when taken in aggregate over the entire 15 week span, the test fails to reject that the two groups are different on average. Still, the effect on the point estimate of the cumulative treatment effect is approximately R$0.06, about 10% of the aggregate treatment effect. Thus, the results of the specification test call into question long run extensions of DID when the pre-period is short and distant from the treatment period as time-varying differences are potentially amplified over time.

C. Countercyclical Treatment Effects?

Before concluding this section, we make one more empirical observation. In Figure 11, we see how the baseline level of sales (measured using total purchases by the control group) and the treatment effect of advertising on sales both vary considerably from week to week. Because the treatment effect is an order of magnitude smaller than the baseline

---

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

level of sales, we use two different vertical scales, as labeled on the right and left of the figure. We discover the surprising result that the treatment effect appears to vary inversely with the level of sales. Figure 12 illustrates this negative relationship with a scatter plot; the best-fit regression line has an R^2 of 0.6.

With higher the level of sales, the lower the absolute magnitude of the treatment effect becomes. As Figure 12 shows, when per-person sales increase (within the sample) from R$0.75 to R$1.75, the treatment effect falls from just over R$0.05 to nearly zero.[16] In other words, the effects of retail advertising look countercyclical: having advertised tends to decrease the variance of weekly sales. This result is rather surprising: for example, if the treatment effect had been a multiplicative percentage of sales, we would have expected the opposite result. Thus, advertising appears to help the retailer boost sales during slower times of the year, smoothing sales.

We do not wish to overstate the representativeness of this result. We note that it comes from just fifteen observations of weekly sales data for a single retailer advertising on one particular website. As such, we consider the result to be merely suggestive rather than definitive. However, we do find the result quite interesting and are unaware of any previous results on this topic. It would be very valuable to see whether this result can be replicated in future advertising experiments, because it has potentially important implications for optimal advertising policies by retailers.

D. 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 huge return to advertising for the period of our sample. It is possible, however, that we are still underestimating the returns to advertising, because we are missing sales data two weeks (near Christmas) between campaigns #2 and #3, and because our sales data end one week after the end of campaign #3, when our previous results indicate that that advertising may well have persistent effects beyond the end of our sample period. We hope to investigate this persistence in future experiments with longer panels of sales data.

VI. Detailed Results for Campaign #1

[16] At first, we worried that this relationship might be a mechanical artifact of the DID estimator, which subtracts control-group sales (as well as the sales of other untreated individuals) from treated individuals' sales, so comparing this difference to the level of control-group sales might be forcing a negative relationship. However, we got a very similar negative slope when we used treatment-group sales as the baseline sales amount, where one might expect mechanical artifacts to force a positive relationship. Having tried both groups to measure baseline sales, we feel sanguine about reporting the correlation using control-group sales.

In this section, we dig deeper into several other dimensions of the data. Data glitches and the relative magnitude of the first campaign relative to the following two make it convenient to examine the effects of the first campaign only. To the extent that any persistence results can be generalized to a longer period of time for the total treatment effect, we suspect similar results may apply to any persistence investigation for the results of this section. We leave the investigation of this interaction between these detailed campaign #1 results and persistence to future research.

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. We perform all of these analyses only on Campaign #1. Because the treatment and control groups were not re-randomized between campaigns and because of the persistence results in the previous section, we know that the second and third campaigns cannot be analyzed cleanly on their own. For the results in this section, we continue to use our preferred specification of a difference in differences for three weeks before and after the start of Campaign #1, comparing treated versus untreated individuals.

A. Offline versus Online Sales and Views versus Clicks

In Table 9, we present a decomposition of the treatment effect into offline and online components by running the previous difference-in-difference analysis separately for offline and online sales. The first line of the table shows that the vast majority of 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.

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 differential impact on sales. An interesting question is, therefore, “To what extent do clicks on ads predict retail sales?”

We answer this question in the second and third lines in Table 9. 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. The results are qualitatively different for offline versus 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 very close to 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 click effect is highly statistically significant for online sales, it is insignificant for online sales due to a large standard error.

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 effect of views versus clicks as follows, 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.

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

We saw in Figure 3 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, we redefine the vertical intercept of the graph so that it equals zero for those with zero ad views.

Figure 13 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 has become so sparse (only 6.1% of the treatment group see more than 100 ad views) that the effect is no longer statistically distinguishable from zero.

We caution that we may not want to interpret this graph as showing the additional sales that would be generated from each marginal increase in 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. On the other hand, we know there is some true exogenous variation in ad views for a given

individual: for example, the ad server has some random component in deciding whether to show this retailer's ad or another advertiser's ad on a given page, and this type of variation in ads would be uncorrelated with a user's type. To the extent that variation in the number of ads represents this kind of within-individual variation rather than variation between different types of users, then a valid causal interpretation would be possible, particularly locally.

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. How the Treatment Effect Varies with Age and Gender

Having seen the significant effect of ad frequency on the treated, we investigate who is actually seeing those ads. Figure 14 shows a nonparametric regression of the average number of ads seen by age group. The number of ads seen is actually highest for the younger age groups between the ages of 25 and 35 with the mean hovering around 45-50 ads per person treated and then declining to an average of approximately 33 for the more elderly. If we believe in an unconditionally increasing marginal effect of an ad, we would expect the treatment effect to be higher for younger age groups than for older groups.

We interact the treatment effect with age and perform a nonparametric regression of first difference sales on age interacted with the treated dummy in Figure 15. As before, we use the Epanechnikov kernel (bandwidth on age of 6.5 years) and a local linear estimate of the conditional mean function. The age distribution of sample participants is most concentrated around 40 years of age, which is where we have the most data and have the most power to be able to identify a significant treatment effect. However, the largest treatment effects of the retailer's advertising appear around age 45 and are large and substantial until age 70, at which point a lack of statistical significance bars further conclusions. So, it would seem that the treatment effect is not correlated with ad

frequency as we might have expected with an unconditionally increasing marginal effect of an ad—age is more important for the effectiveness of the ads than just frequency.

When the treatment effect is decomposed by gender, there is an insignificant, difference. Men are slightly more affected by the advertising campaign than women, with the three-week campaign's DID estimate being R$0.201 (0.066) for men and R$0.141 (0.060) for women. However, when we examine the nonparametric plots of gender interacted with age in Figure 16 and Figure 17, we see that there is somewhat different behavior between the two estimates. The level of pointwise significance in the difference between the two has not been determined, but the point estimates are still of interest. Women are suddenly highly affected for ages between 50 and 70, while men, which compose only 40.4% of the sample, are more gradually and less significantly affected starting around age 40 while increasing until age 70. Further, in terms of per ad average effectiveness, for all age groups women on average view significantly fewer ads (35.5 v. 45.6 ads with p=0.000 for the difference). Thus, for those age groups where there is a large and significant treatment effect, the advertising is much more cost-effective.

The nonparametric estimates of the treatment effect interacted with age and gender show significant differences in the effectiveness of the advertising for different age groups and suggestive, although not stark, differences between men and women.

D. 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 10 for comparison to our treatment-effect results.

In Table 10, 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,[17] 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).

[17] 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.

This represents an increase of approximately 1.6% relative to the average probability of a transaction.

Next, we consider the effect on basket size. Instead of computing this DID estimate via regression using first-differenced sales as before, we use group means of basket size, paying careful attention to possible time-series correlation in order to compute a consistent standard error.[18] As shown in the second row of Table 10, the advertising campaign produced an increase in basket size of 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. 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 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! The randomized experiment allows us to establish causal effects of advertising on sales. Panel data on weekly individual transactions allows us to measure these effects rather precisely, despite the fact that sales at this retailer have high variance and this online advertising campaign is just one of many factors that influence purchases. The panel nature of the data also illuminates how we would make vastly incorrect measurements of the effect of advertising on sales if we attempted to estimate these effects using non-experimental cross-section variation in advertising exposure.

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

Though it may be tempting to assume that online advertising would have disproportionately large effects on online retail sales, we find the reverse to be true. This

[18] 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.

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 those who merely view the ads than on those who actually click them. Clicks are a good predictor of online sales, but not for 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. Finally, we find evidence suggesting that the effects of having advertised to customers may be countercyclical, in the sense that advertising's impact is greatest in weeks when total sales are smallest.

In future research, we hope to replicate these results with other retailers. 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. We believe John Wanamaker would be very interested in the results from this new frontier.

References

Abraham, Magid. "The Off-Line Impact of Online Ads." *Harvard Business Review*, April 2008, p. 28.

Abraham, Magid, and Len Lodish. "Getting the Most out of Advertising and Promotion." *Harvard Business Review*, May-June 1990, pp. 50-60.

Ackerberg, Daniel. "Empirically Distinguishing Informative and Prestige Effects of Advertising." *RAND Journal of Economics*, vol. 32, no. 2, Summer 2001, pp. 316-333.

Ackerberg, Daniel. "Advertising, Learning, and Consumer Choice in Experience-Good Markets: An Empirical Examination." *International Economic Review*, vol. 44, no. 3, August 2003, pp. 1007-1040.

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

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

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

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

Dorfman, Robert, and Peter O. Steiner. "Optimal Advertising and Product Quality." *American Economic Review*, vol. 44, no. 5, December 1954, 826-836.

Hu, Ye, Leonard M. Lodish, and Abba M. Krieger. "An Analysis of Real World TV Advertising Tests: a 15-Year Update." *Journal of Advertising Research*, vol. 47, no. 3, September 2007, pp. 341-353.

Levitt, Steven, and John A. List. "Field Experiments in Economics: The Past, the Present, and the Future." NBER Working Paper 14356, 2008.

Lewis, Randall. "Data Recovery: Least Squares Consistency for Inexact Merges." Working paper, MIT, 2008.

Lodish, Leonard M., Magid Abraham, Stuart Kalmenson, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens. "How T.V. Advertising Works: a Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments." Journal of Marketing Research, vol. XXXII, May 1995a, pp. 125-139.

Lodish, Leonard M., Magid Abraham, Stuart Kalmenson, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens. “A Summary of Fifty-Five In-Market Experiments of the Long-Term Effect of TV Advertising.” *Marketing Science*, vol. 14, no. 3, part 2, 1995, pp. G133-140.

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

Tables and Figures

Table 1 – Summary Statistics for the Three Campaigns

	Campaign 1	Campaign 2	Campaign 3	All 3 Campaigns
Time Period Covered	Early Fall '07	Late Fall '07	Winter '08	
Length of Campaign	14 days	10 days	10 days	
Number of Ads Displayed	32,272,816	9,664,332	17,010,502	58,947,650
Number of Users Shown Ads	814,052	721,378	801,174	924,484
% Treatment Group Viewing Ads	63.7%	56.5%	62.7%	72.3%
Mean Ad Views per Viewer	39.6	13.4	21.2	63.8

Figure 1 – Time Plot of Treatment Saturation

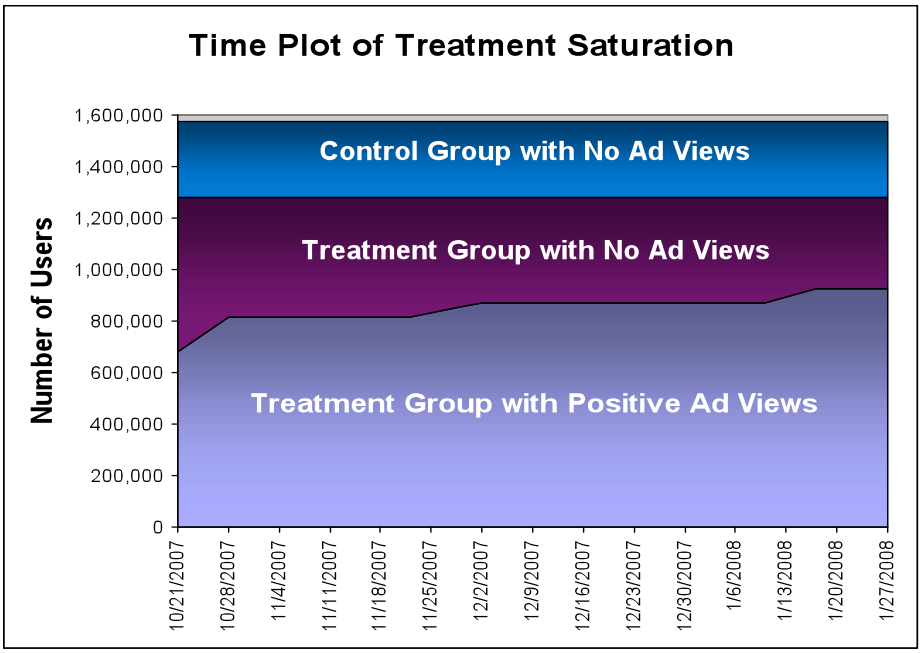

Figure 2– Yahoo! Front Page with Large Rectangular Advertisement

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%
% People Clicking at Least Once	-	4.59%

Figure 3 - Ad Views Histogram

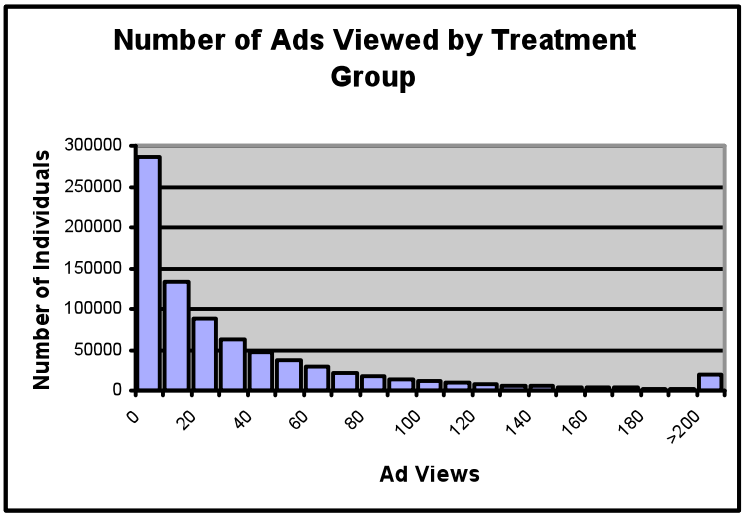

Table 3 – Weekly Sales Summary

		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
Campaign #3						
12/21	3 Weeks Before	**R$ 1.635**	17.9	-2051.39	2521.88	62,454
12/28	2 Weeks Before	**R$ 0.812**	13.0	-1238.83	1870.99	49,144
01/04	1 Week Before	**R$ 0.616**	11.7	-1120.77	3400.54	38,265
01/11	Week 1 During	**R$ 0.644**	11.7	-1118.58	3939.81	36,321
01/18	Week 2 During (3 Days)	**R$ 0.322**	7.5	-588.84	1437.17	18,238
01/21	Week 1 Following	**R$ 0.636**	11.5	-2336.83	3300.97	33,724

N=1,577,256 observations per week

Figure 4 - Offline and Online Weekly Sales

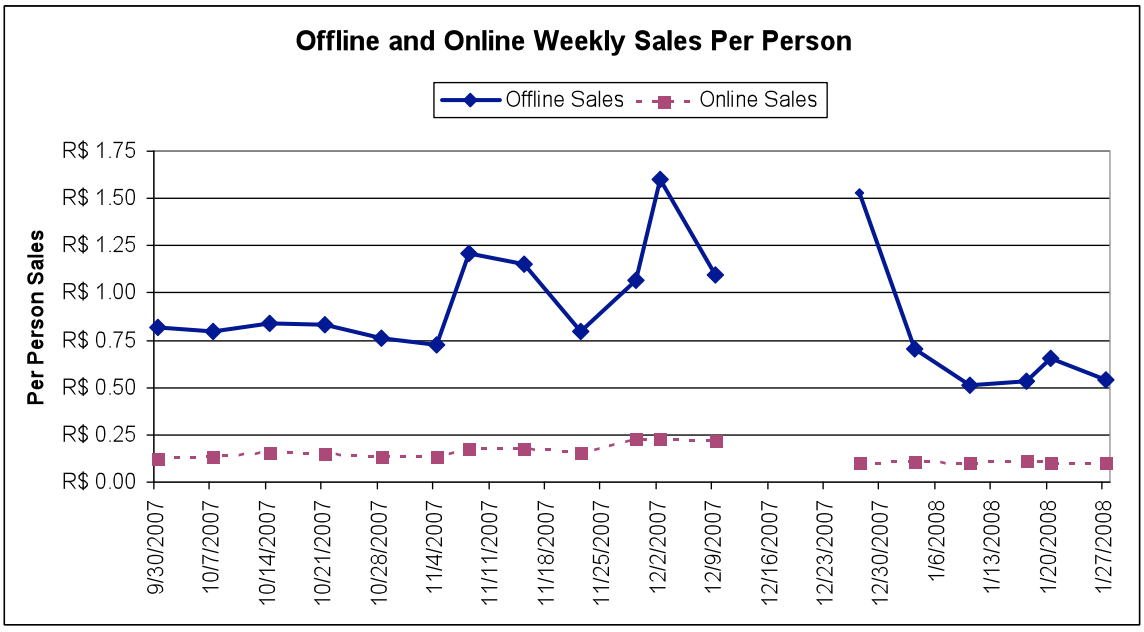

Table 4 - Three Week Treatment Effect Offline/Online Decomposition

	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 5 - Histogram of Campaign #1 Sales by Treatment and Control

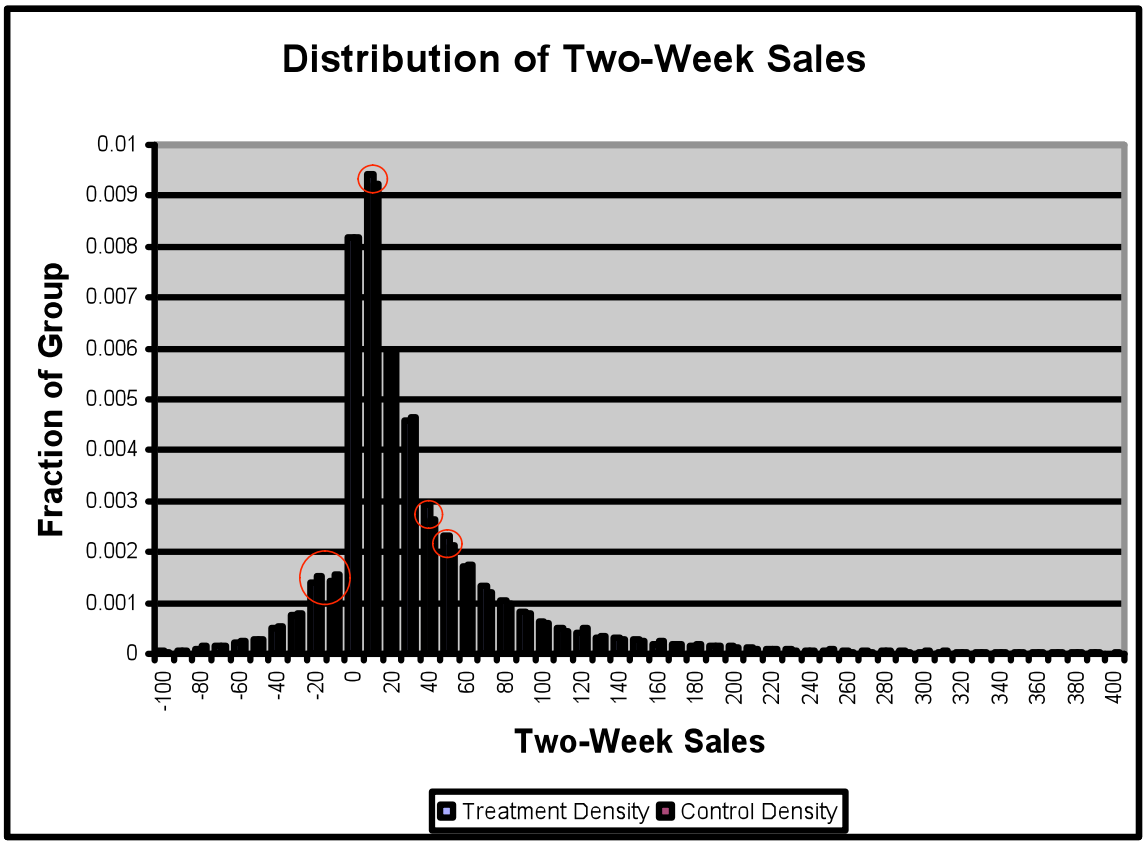

Figure 6 - Difference between Treatment and Control Sales Histograms

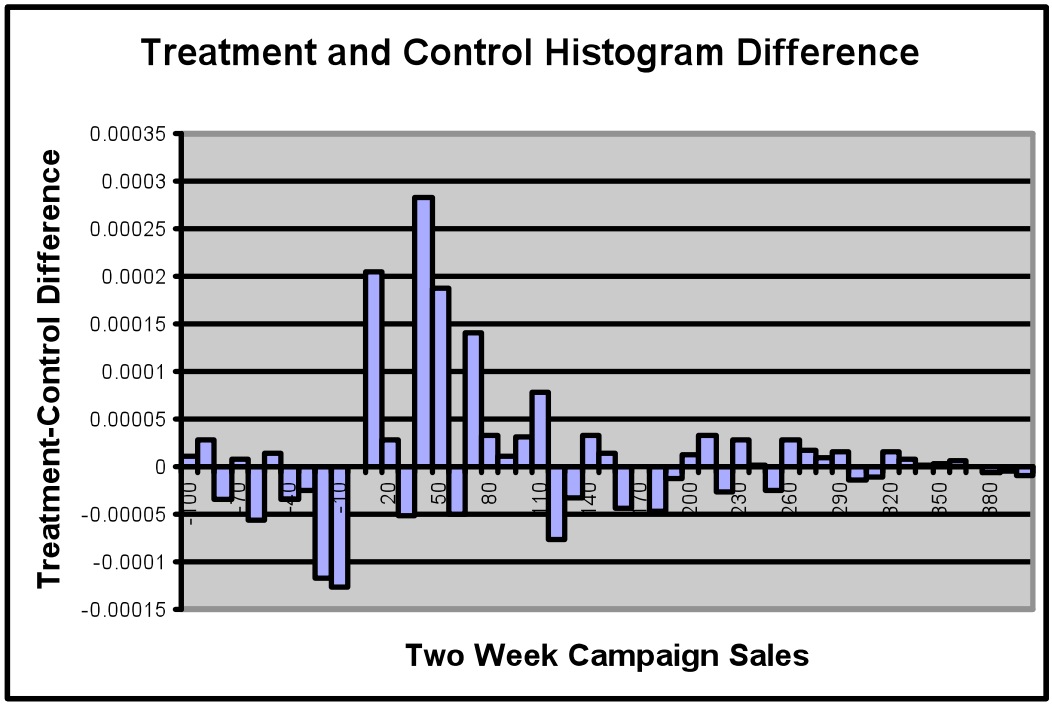

Table 5 - Preliminary Two-Week Treatment Effect Estimates

	All Individuals	Excluding Page Views=0
Treatment-Control Difference	R$ 0.053	**R$ 0.078**
	(0.038)	(0.045)
Rescaled Effect on Treated	0.083	**0.093**
	(0.059)	(0.054)

Figure 7 – Histogram of Difference in Three-Week Sales for Treated and Untreated Groups

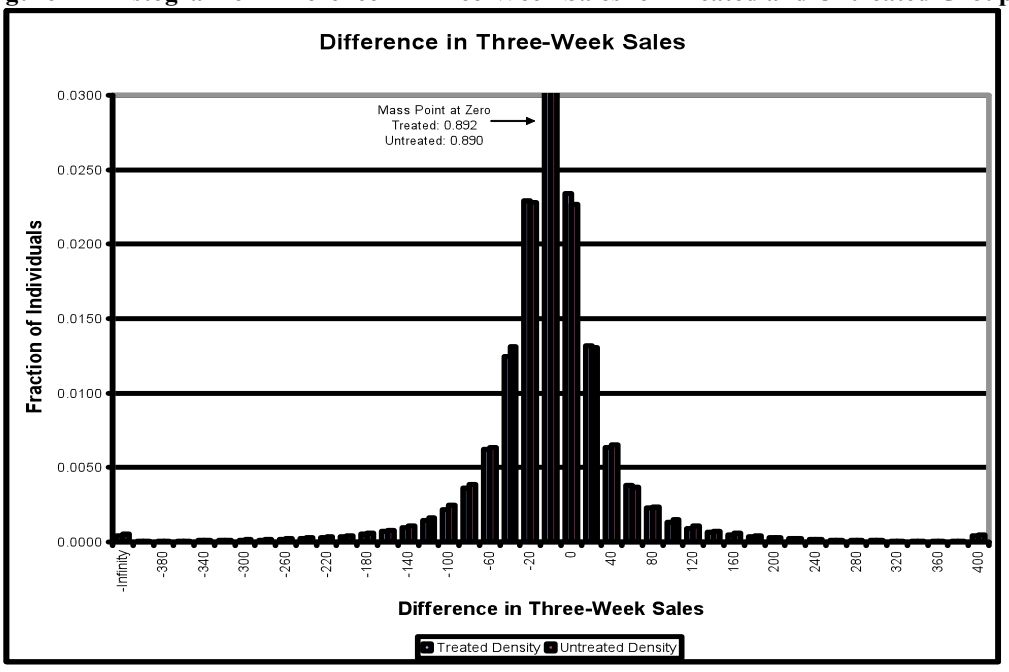

Figure 8 – Difference in Treated and Untreated Three-Week Sales Histograms

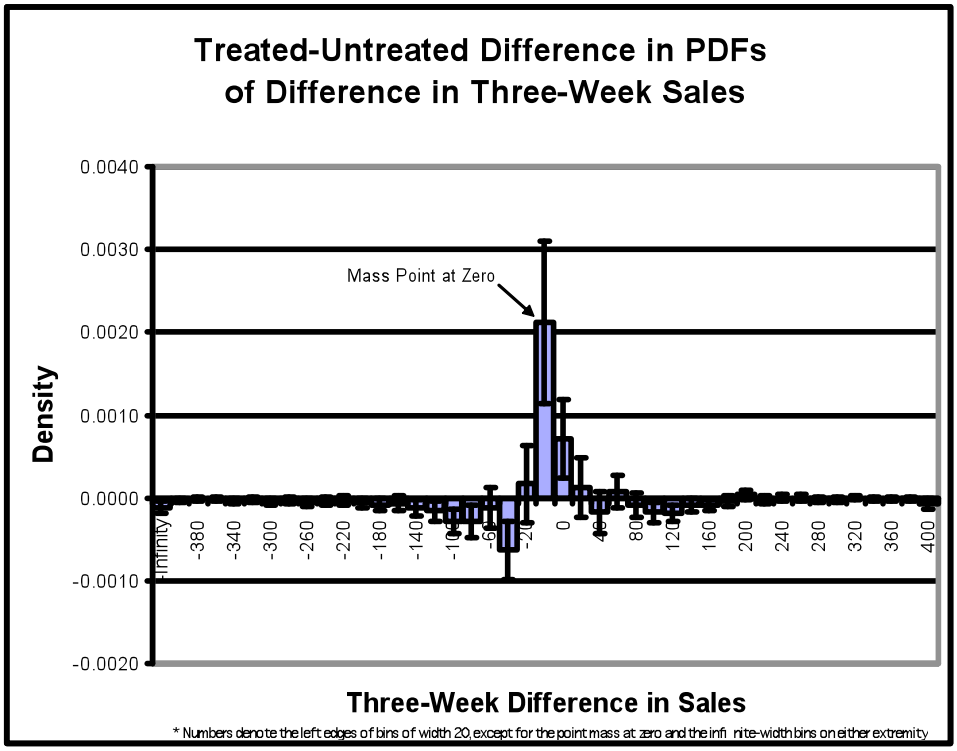

Table 6 – Difference-in-Differences Treatment Effect Estimate

	Two-Week	Three-Week
Difference in Difference	**R$ 0.102**	**R$ 0.166**
	(0.043)	(0.052)

Table 7 – 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
Campaign #3**		
3 Weeks Before	**R$ 0.029**	0.032
2 Weeks Before	**R$ 0.060**	0.025
1 Week Before	**R$ 0.064**	0.023
Week 1 During	**R$ 0.080**	0.023
Week 2 During (3 Days)	**R$ 0.035**	0.013
Week 1 Following	**R$ 0.049**	0.023
	N=1,577,256 obs. per week**	

* For purposes of computing the treatment effect on the treated, we define "treated" individuals as having ever seen an ad in one of these campaigns up to that point in time.

** Estimates for Campaign #3 involve an imperfect merge to compute the difference in differences. See footnote 14.

Figure 9 - Three Campaign Weekly Treatment Effect Time Plot

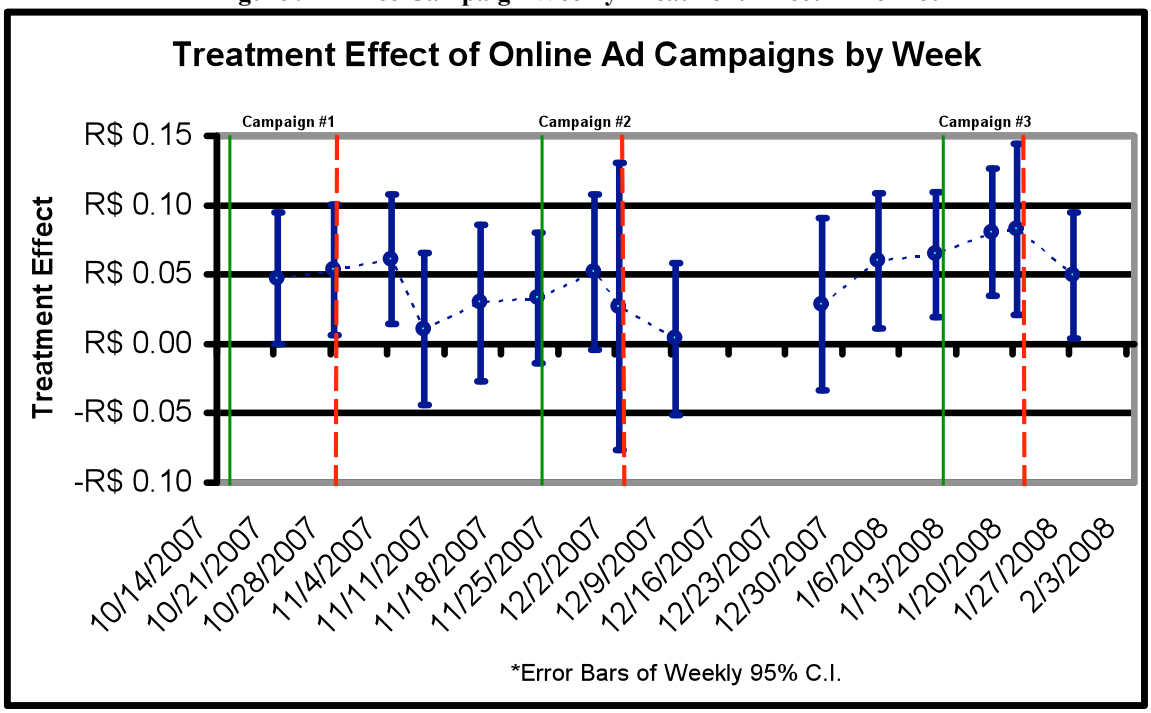

Table 8 - Three Campaign Results Summary

	Treatment Effect	Robust S.E.	t-stat	P(|t|>T)
Average Weekly Effect				
Simple Average (OLS)	**R$ 0.045**	0.0140	3.25	0.001
Efficient Average (GLS)	**R$ 0.048**	0.0136	3.53	0.000
Cumulative Effects over All 3 Campaigns				
Cumulative Sales	**R$ 0.532**	0.196	2.72	0.007
Simple Aggregate Effect (OLS)	**R$ 0.611**	0.188	3.25	0.001
Efficient Aggregate Effect (GLS)	**R$ 0.645**	0.183	3.53	0.000
Length of Measured Cumulative Effects		13 wks. 3 days		

Figure 10 - Weekly DID Specification Test

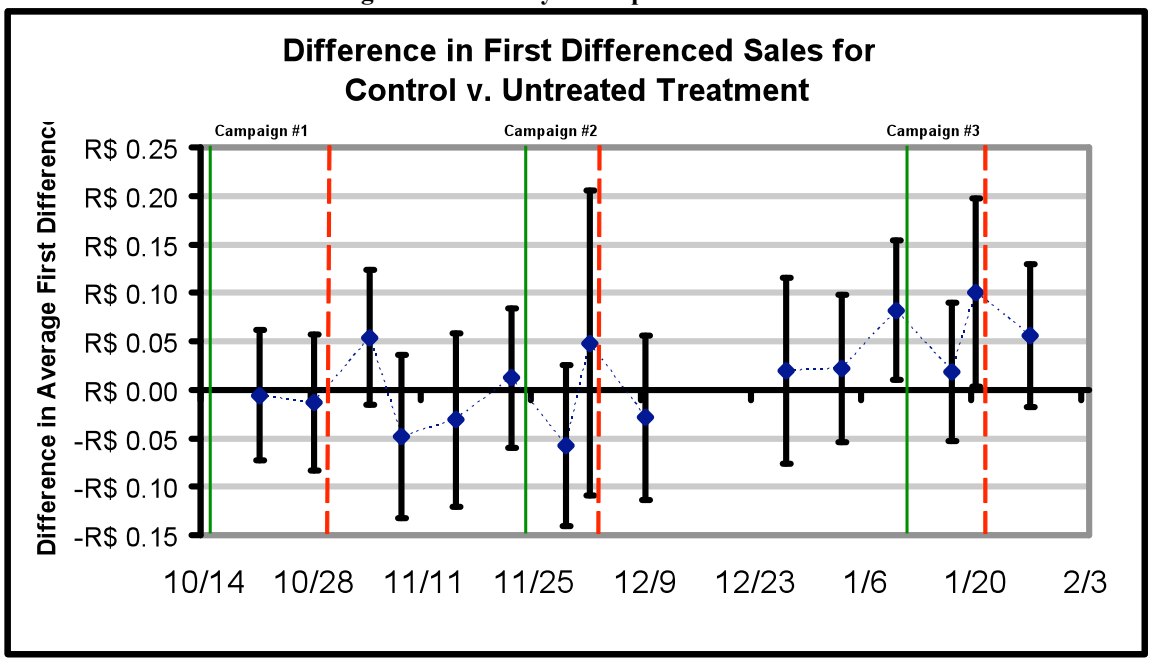

Figure 11 - Total Sales Histogram with Treatment Effect Estimates

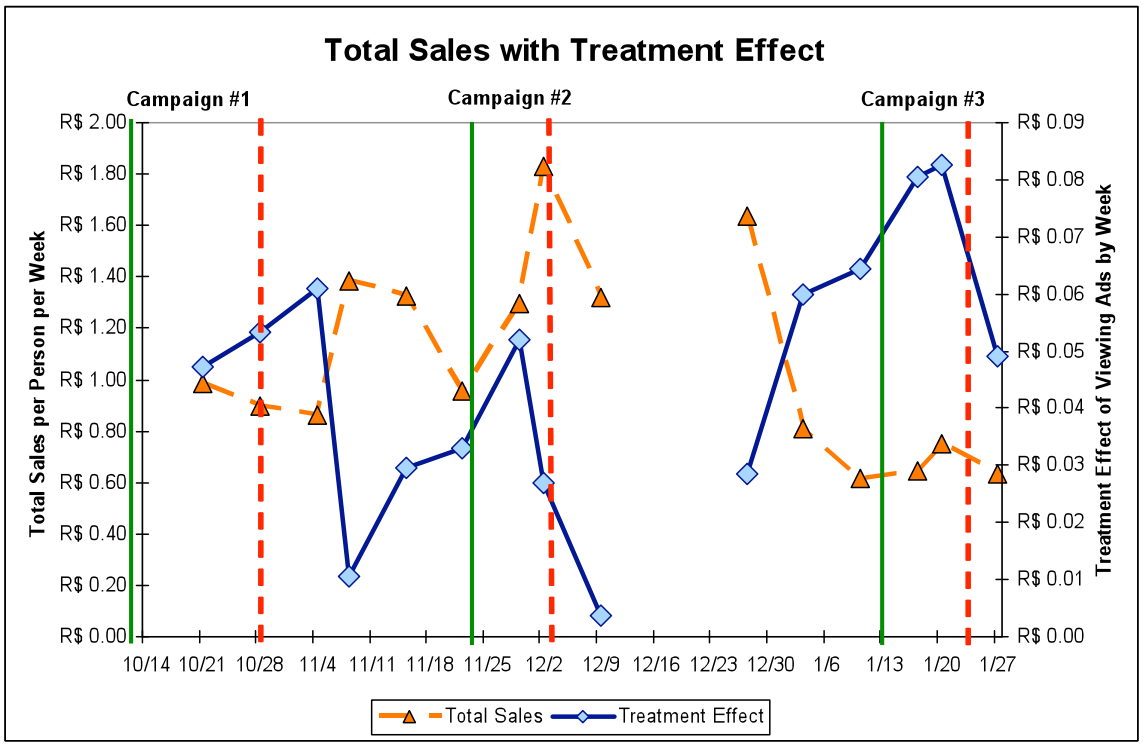

Figure 12 - Plot and Regressions of Weekly Treatment Effect with Sales

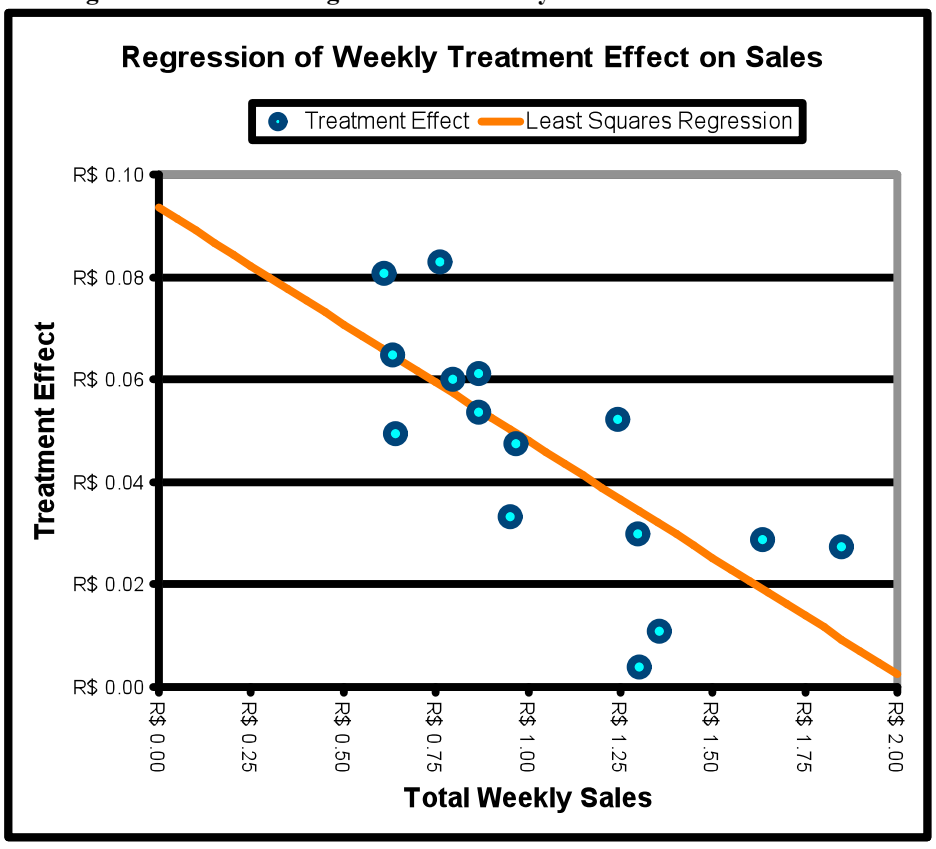

Table 9 - Offline/Online Treatment Effect Decomposition

	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 13 – Nonparametric Estimate of the Treatment Effect by Ad Viewing Outcome

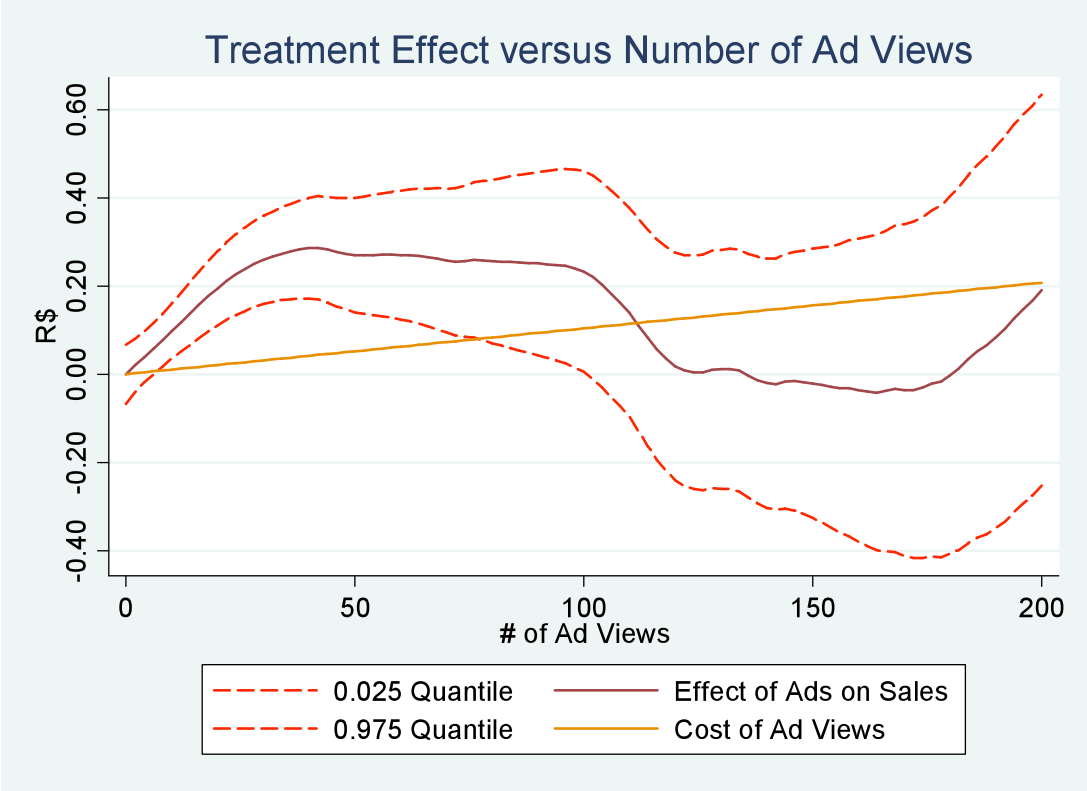

Figure 14 - Nonparametric Estimate of the Average Campaign #1 Exposure by Age

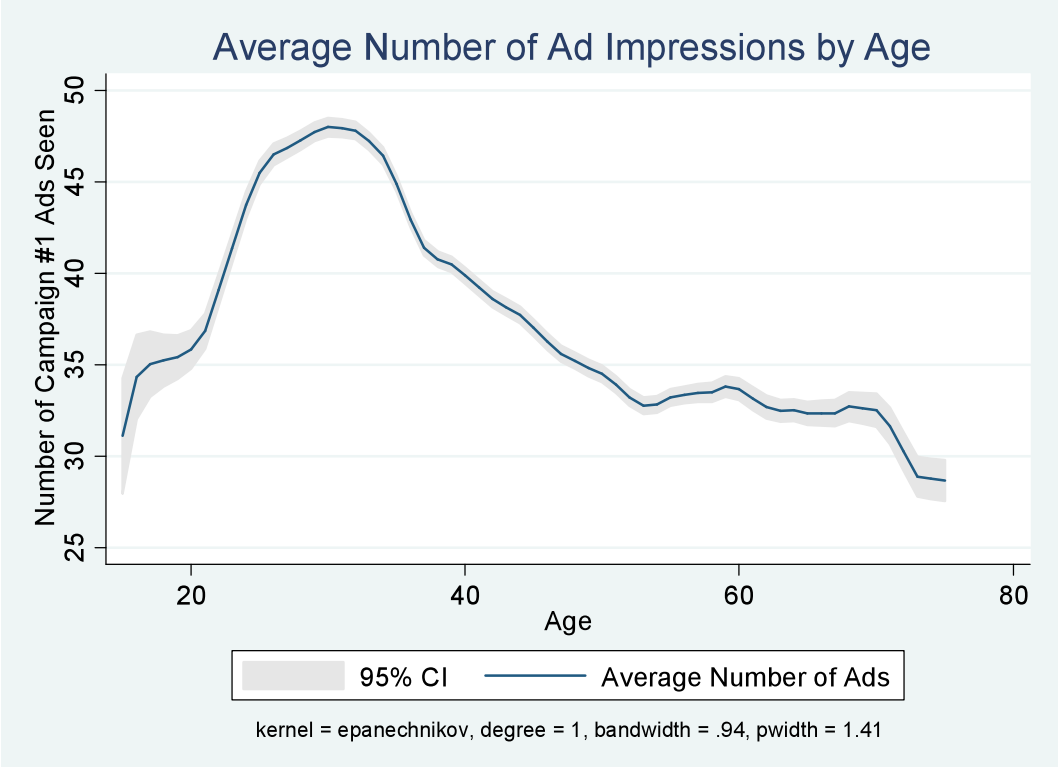

Figure 15 - Nonparametric Estimate of the Treatment Effect by Age

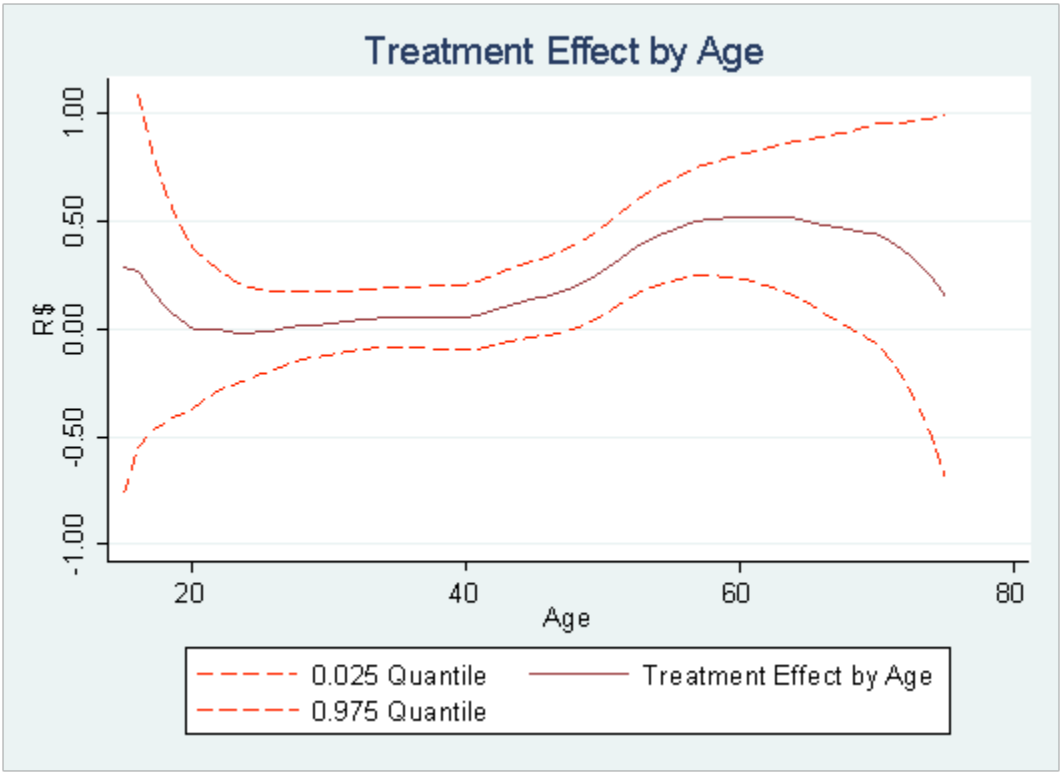

Figure 16 - Nonparametric Estimate of the Treatment Effect for Females

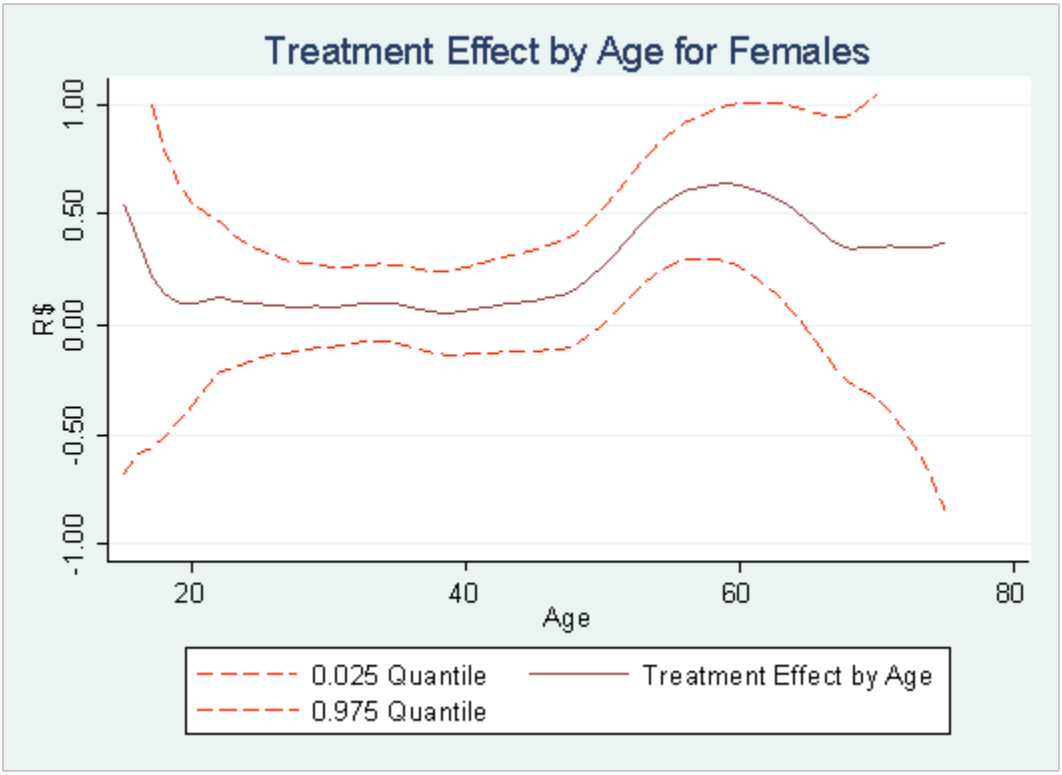

Figure 17 - Nonparametric Estimate of the Treatment Effect for Males

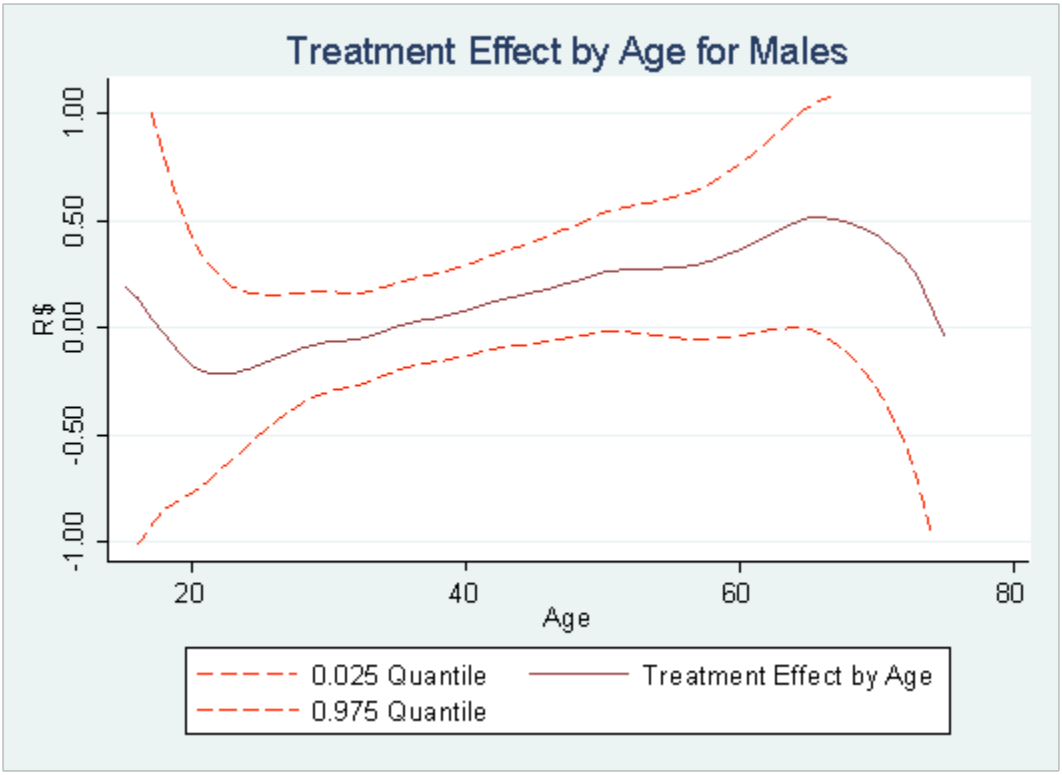

Table 10 - Decomposition of Treatment Effect into Basket Size and Frequency Effects

	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.