Quant Mark Econ (2014) 12:235–266
DOI 10.1007/s11129-014-9146-6

Online ads and offline sales: Measuring the effects of retail advertising via a controlled experiment on Yahoo!

Randall A. Lewis · David H. Reiley

Received: 1 October 2013 / Accepted: 28 March 2014 / Published online: 28 May 2014

Abstract A randomized experiment with 1.6 million customers measures positive causal effects of online advertising for a major retailer. The advertising profitably increases purchases by 5 %. 93 % of the increase occurs in brick-and-mortar stores; 78 % of the increase derives from consumers who never click the ads. Our large sample reaches the statistical frontier for measuring economically relevant effects. We improve econometric efficiency by supplementing our experimental variation with non-experimental variation caused by consumer browsing behavior. Our experiment provides a specification check for observational difference-in-differences and cross-sectional estimators; the latter exhibits a large negative bias three times the estimated experimental effect.

Keywords Online advertising · Display advertising · Advertising effectiveness · Field experiment · Difference in differences

JEL Classifications C93 · M37 · D12

This work was completed while both authors were employees at Yahoo! Research. Previously circulated versions were titled "Does Retail Advertising Work?" and "Retail Advertising Works!"

R. A. Lewis · D. H. Reiley (✉)
Google, Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
e-mail: david@davidreiley.com

R. A. Lewis
e-mail: randall@econinformatics.com

Online advertising now represents 21 % of all advertising in the United States, but online retail represents only 5.5 % of all retail purchases.[1] This divergence makes sense if online advertising's effects go beyond e-commerce to offline purchases. However, the offline impact has been difficult to measure. In this paper, we make a significant advance in such measurement, reporting the results of a controlled experiment with an unprecedented number of individuals, randomizing exposure to online advertising, and linking it to data on both offline and online purchases.

At more than 1 % of GDP, the hundreds of billions of dollars spent on advertising each year[2] not only represent substantial economic activity but also raise many interesting economic questions: how does advertising change consumer behavior, affect firm competition, and impact consumer welfare? However, before we can credibly obtain answers to these deep questions, we first need an empirical strategy to identify the causal effects of advertising. Our research takes a major step in the direction of understanding the effects of brand advertising on consumer purchases.

The most unique feature of this research is our ability to demonstrate the effectiveness of online advertising on in-store purchases. In general, online advertising has provided greater measurement opportunities than has advertising in traditional media. Online advertisers get one automatic measurement of interest through reports on the number of clicks on the ads. With additional effort, they can also install software to track the number of online "conversions" (e.g., making a purchase, filling out a form) that consumers make after viewing or clicking a particular ad. However, the focus on online measurement has not created many opportunities to study effects on brick-and-mortar purchases. This research provides a unique opportunity to do so through the matching of individual purchase data at a retail store to individual user accounts on Yahoo!

Unlike most studies of advertising, which typically study advertising by manufacturers, we study the effects of a retail image advertising campaign. Our advertising aims to generate additional purchases for a retailer by targeting its previous customers with advertising that promotes a positive image of the store. Instead of featuring a special offer, price discount, or product details, this advertising simply features beautiful images of products and emphasizes the name of the retail store.

Brand image advertising presents difficult measurement problems because of the indirect nature of its effects. Direct-response advertising (such as catalog mailings, "call now" TV ads, and most search advertising) measures its success by immediate responses. Online display ads do generate some clicks leading to the online store, but a brand image campaign is designed to produce longer-term consumer goodwill as well. In this paper, we provide evidence that online ads do produce effects beyond the click—on purchases both online and in physical stores. We also show that the majority of the impact on purchases comes from viewers who never clicked the ads.

[1] Online is $36 billion relative to approximately $176 billion, or roughly 21 % (IAB, "2012 Annual Report," http://www.iab.net/about_the_iab/annual_report); 5.5 % of all retail purchase are done online (US Census Bureau, "Quarterly Retail e-Commerce Sales: 1st Quarter 2013," http://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf).

[2] $176 billion is spent on advertising according to the 2012 IAB report; 2012 US GDP was $15.7 trillion according to the US Census Bureau. See footnote 1 for more details.

Controlled experiments quantifying the effects of brand advertising on sales are rather rare in practice.[3] 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. Indeed, the relationship between sales and advertising is a textbook example of the endogeneity problem in econometrics: Berndt (1991) reviews the observational literature on advertising in his applied-econometrics text, which emphasizes the simultaneity problem in firms' advertising choices.

As Levitt and List (2009) point out, field experiments have become increasingly important in economics and the social sciences as researchers have recognized their value for creating exogenous variation and eliminating the econometric problems of selection and omitted-variable bias. The field of marketing produced a number of early field experiments on the effects of brand advertising. As we briefly review this literature, we will note a feature that we rediscovered in our own work: statistical precision is quite difficult to obtain in measuring the effects of brand advertising on purchases.

Early marketing experiments generated randomized advertising exposure across a small number of geographic units. Ackoff and Emshoff (1975) evaluated the effects of increased or decreased quantities of advertising on Budweiser beer sales with six geographical areas per treatment.[4] Eastlack and Rao (1989) reported on a series of advertising experiments for Campbell's Soup with 31 geographic units of observation. Aaker and Carman (1982, pp. 59-61) review 69 early experiments reported in nine different sources that vary the level of advertising; fewer than 40 % of the individual tests produced statistically significant results, despite using mild statistical significance thresholds (at least half of the experiments used significance levels of 0.40). One could read these results as saying that sometimes advertising works and sometimes it does not, but we have come to interpret the absence of statistical significance as symptomatic of low statistical power.

The most significant predecessors to our work are the studies exploiting IRI's BehaviorScan technology, specifically developed for advertisers to experiment with television ads and measure the effects on sales. These studies split a cable-TV signal to generate more exposure for a given television ad in treatment versus control groups. They measured supermarket sales via scanner data for panels of approximately 3,000 households whose TV exposure could be manipulated. 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

[3] By contrast, experiments are more common in direct-response advertising. Direct-mail advertisers have a culture of randomizing aspects of the mailings—even minute details such as ink color and envelope size. We were fortunate to have a partner at our retailer who previously worked in retail catalog mailings and was therefore familiar with the benefits of experimentation.

[4] Allaire (1975) pointed out that the authors had failed to quantify the uncertainty in their estimates, and their interesting effects turned out to be statistically insignificant.

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) investigate long-run effects, showing that for the subset of ads that produced statistically significant results during a year-long experiment, positive effects also tend to obtain in the two subsequent years. Hu et al. (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.

We began this research project expecting that an experiment with more than a million customers would give precise statistical results—we now think otherwise. As we will demonstrate, an economically significant (i.e., profitable) effect of advertising 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 is like looking for a needle in a haystack.[5] As a result, we understand why the experimental studies reviewed above produced point estimates that were usually statistically insignificant at conventional levels. By studying over a million users, we shrink confidence intervals to the point where effects of economically interesting magnitudes have a reasonable chance of being statistically significant. Even then, to obtain sharp statistical significance we choose to augment our experimental variation with potentially endogenous non-experimental variation in the data.

For an example of an observational study of the economics of advertising, we consider the work of Ackerberg (2001, 2003).[6] This work uses panel-data methods on individual-level matched data on yogurt advertising and purchases for 2,000 households. By exploiting the panel nature of the dataset for credible identification, Ackerberg shows positive effects of advertising for a new product (Yoplait 150), particularly for consumers previously inexperienced with the product.

Because our data, like Ackerberg's, has a panel structure with individual sales data both before and after the advertising campaign, we employ a differences-in-differences (DID) estimator[7] that exploits both experimental and non-experimental variation in ad exposure. DID yields a similar point estimate to the simple experimental difference, but with higher precision. We therefore prefer the more efficient

[5] By contrast, direct-response advertising may produce more statistical power in experiments than brand advertising, because the ads are more salient (higher signal) and produce more immediate responses (less noise). This may explain why direct-mail marketers are more likely to engage in experimentation than other advertisers (see footnote 3). Recent examples in the academic literature include Simester et al. (2009), who experimentally vary the frequency of catalog mailings; Bertrand et al. (2010), who vary the ad creative and the interest rate in loan offers by mail; and Ghose and Yang (2009), who measure the impact of sponsored search advertisements on total clicks for the advertiser on the search-results page.

[6] For a survey of empirical and theoretical work on the economics of advertising, see Bagwell (2008). DellaVigna and Gentzkow (2010) reviews empirical work on persuasion more generally, including advertising and other communication to charitable donors, voters, investors, and retail consumers.

[7] Other panel models satisfy a similar role; we use DID for its conceptual simplicity.

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). In principle, our preferred estimator could have been computed in the absence of an experiment, but 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 variation in advertising exposure, thus increasing the precision of the estimate.

The paper proceeds as follows. We present the design of the experiment in Section 1 and describe the data in Section 2. In Section 3, we measure the effect on sales during the experimental advertising campaign. In Section 4, we measure the persistence of advertising's effect after the end of the campaign and after a much smaller follow-up campaign, asking whether the ads produce incremental sales or merely accelerate consumer purchases forward in time. In Section 5, we examine how the treatment effect of online advertising varies across several dimensions that we hope will be useful in developing the theory of advertising. These include the effect on online versus offline sales, the effect on users who click ads versus those who merely view, the effect on users who see few versus many ads, and the effect on consumers' probability of purchasing versus average purchase size. The final section concludes.

1 Experimental design

This experiment randomized individual-level exposure to a nationwide retailer's display-advertising campaign on Yahoo! To measure the causal effects of the advertising on individuals' weekly purchases both online and in stores, we matched the retailer's customer database against Yahoo!'s user database, yielding 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 the retailer's existing customers who log in to Yahoo![8] Of these matched users, we randomly assigned 81 % to a treatment group who were eligible to see ads from the retailer's two campaigns on Yahoo! The remaining 19 % served as a control group who saw none of the retailer's ads on Yahoo! The simple randomization ensures the treatment-control assignment is independent of all other variables.

We first ran a two-week advertising campaign in the fall of 2007 in the treatment group. Several weeks later we ran a follow-up advertising campaign, without re-randomizing, so the treatment group remained constant for both campaigns. We will focus most of our analysis on the first campaign (hereafter, "the campaign") because it is much larger, comprising 77 % of both campaigns' ad views. Because persistent effects of the campaign may prevent us from separately identifying the effects of the

[8]The retailer selected for the match a subset of their customers to whom they wished to advertise. We do not have precise information about their selection rule.

Table 1 Summary statistics for the campaigns

	The campaign	Follow-up campaign	Both
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 viewing ads	63.7 %	56.5 %	67.9 %
Mean ad views per viewer	39.6	13.4	48.3

follow-up campaign, we examine the latter only in Section 4 on long-run effects of the advertising.[9]

Table 1 gives summary statistics for the campaign and the follow-up campaign, which delivered 32 million and 10 million impressions, respectively.[10] The campaign exposed 814,000 users to advertising for the retailer, while the follow-up campaign increased the total number of exposed users to 868,000. Exposed individuals received an average of 40 ad impressions per person in the first campaign, versus 48 ad impressions across both campaigns. Control-group users, by contrast, saw a variety of other advertisements: they saw whatever ads would have been served to the treatment group if the retailer had not purchased this campaign on Yahoo! This makes the control-treatment difference exactly the right comparison to answer the question, "What is the total difference in sales caused by this retailer's ad campaign?"

These were the only ads shown by this retailer on Yahoo! during this time period. However, Yahoo! ads represent only a fraction of the retailer's overall advertising budget, which included other media such as newspaper and direct mail. Fortunately, randomization makes the Yahoo! advertising uncorrelated with any other influences on shopping behavior, including other ad campaigns on other media. Our experimental estimate will therefore give an unbiased estimate of the causal effects of the Yahoo! advertising. By contrast, our preferred DID estimate could, in principle, be biased by simultaneous advertising in other media affecting our treated and untreated consumers differently; we discuss this possibility in more detail below.

The ads were "run-of-network" ads on Yahoo! This means that ads appeared on various Yahoo! sites, such as mail.yahoo.com, groups.yahoo.com, and maps.yahoo.com. Figure 1 shows a typical display advertisement placed on Yahoo!

[9]Early drafts of this paper also examined a third campaign, whose analysis required an imperfect data merge. For improved data reliability and simplicity of exposition we now choose to omit all references to the third campaign.

[10]The industry uses "impressions" as the standard accounting unit to refer to "online ads that loaded in a web page on the recipient's computer." While there is no guarantee that a given impression rendered on a user's screen or received any visual attention by the internet user, for simplicity of exposition and in accordance with industry practice we use the words "impression," "exposure," and "view" synonymously.

Fig. 1 Yahoo! Front page with large rectangular advertisement

This large rectangular ad for Dell[11] 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 so that neither party could identify individual users in the matched dataset. In addition, the retailer disguised actual sales amounts by multiplying all sales figures by an undisclosed number between 0.1 and 10, constant across observations. All financial quantities in this paper, including sales and costs, will be reported in these rescaled "Retail Dollars" (R$) rather than US dollars.

2 Data description

We describe the Yahoo! advertising data and the retailer sales data for the set of matched customers. Then, after highlighting the high variance of the sales data, we discuss the statistical power of our experiment to detect economically meaningful treatment effects due to advertising.

[11] Dell was not the retailer in this experiment—the retailer prefers anonymity.

Table 2 Basic summary statistics for the campaign

	Control	Treatment
% Female	59.50 %	59.70 %
% Retailer ad views > 0	0.00 %	63.70 %
% Yahoo page views > 0	76.40 %	76.40 %
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	-	7.20 %

2.1 Advertising data

Table 2 shows summary statistics consistent with a successfully randomized experiment.[12] 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 two-week 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 these ads on 7.0 % of the pages she visited on Yahoo! 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.

Among Yahoo! users who saw one or more of the campaign's ads, the number of ad views is quite skewed, as shown in Fig. 2. The 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 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 extreme browsing behavior. Most users saw fewer than 100 ads, with a mere 1.0 % viewing more than 500 ads during the two-week campaign. The largest number of ad views by a single user was 6,050.[13]

[12] Only one difference between treatment groups in this table is statistically significant. The mean number of Yahoo! page views was 363 for the treatment group versus 358 for the control, a statistically significant difference ($p = 0.0016$). This difference is rather small in magnitude and largely driven by outliers: 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, remove all bot-like users 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.

[13] 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 our dataset were charged to the retailer as valid impressions.

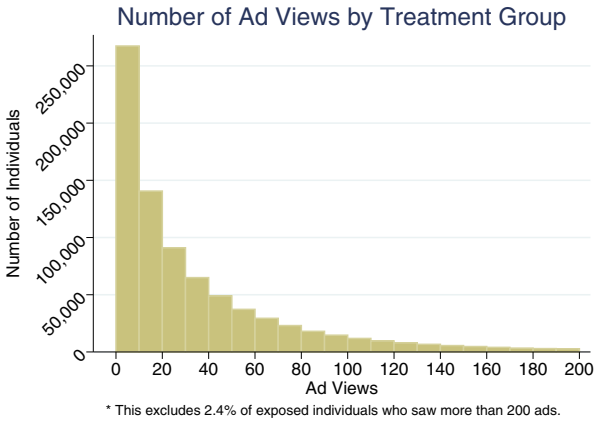

Fig. 2 Ad vews histogram for the campaign

One standard measure of online ad effectiveness is the click-through rate (CTR): the fraction of ads clicked. Table 2 shows that the campaign's CTR was 0.28 %. Our individual data also allow us to calculate the "clicker rate": 7.2 % of those who received ads clicked at least one of them. Below, we will give evidence that ads affect not only the users who click the ads but also those who do not.

2.2 Sales data

Table 3 provides a weekly summary of the sales data, spanning approximately twelve weeks. We have three weeks preceding, two weeks during, and one week following both the campaign and follow-up campaign. Sales amounts include all purchases that the retailer could link to each individual customer in the database.[14] Total sales is obtained by combining online and offline (in-store) sales representing 15 % and 85 % of the total.

Table 3 shows that in spite of being averaged across 1.6 million individuals, mean weekly sales are rather volatile, ranging from R$0.86 to more than R$1.39 per person. Holidays, store-wide promotions, and other seasonal events contribute to these wide swings in sales. Even higher variance can be seen across individuals within weeks, with a standard deviation of approximately R$14 each week. The mean includes a large mass of zeroes, as fewer than 60,000 (4 %) individuals transact with the store in any given week. For those who do transact, the purchases exhibit large positive and negative amounts, though well over 90 % of purchase amounts lie between -R$100 and +R$200. Negative purchase amounts represent net returns of merchandise; we

[14] If these customers make purchases that cannot be tracked by the retailer, our estimate will underestimate the total effect of advertising on sales. However, the retailer claims to attribute 90 % of purchases to the correct customer account via several methods, such as matching the name on a customer's credit card at checkout.

Table 3 Weekly sales summary

		μ	σ	Min	Max	Customers	Online
The Campaign							
09/24	3 Weeks Before	R$ 0.939	14.1	-932.04	4156.01	42,809	13.2 %
10/01	2 Weeks Before	R$ 0.937	14.1	-1380.97	3732.03	41,635	14.8 %
10/08	1 Week Before	R$ 0.999	14.3	-1332.04	3379.61	43,769	16.0 %
10/15	Week 1 During	R$ 0.987	13.5	-2330.10	2163.11	43,956	15.4 %
10/22	Week 2 During	R$ 0.898	13.3	-1520.39	2796.12	40,971	15.4 %
10/29	Week 1 Following	R$ 0.861	13.3	-1097.96	3516.51	40,152	15.7 %
Follow-up Campaign							
11/02	3 Weeks Before	R$ 1.386	16.4	-1574.95	3217.30	52,776	12.9 %
11/09	2 Weeks Before	R$ 1.327	16.6	-654.70	5433.00	57,192	13.4 %
11/16	1 Week Before	R$ 0.956	13.4	-2349.61	2506.57	45,359	16.4 %
11/23	Week 1 During	R$ 1.299	16.7	-1077.83	3671.75	53,428	17.6 %
11/30	Week 2 (3 Days)	R$ 0.784	14.0	-849.51	3669.13	29,927	12.4 %
12/03	Week 1 Following	R$ 1.317	16.1	-2670.87	5273.86	57,522	16.7 %

N=1,577,256 observations per week
"Customers" refers to the number of transacting customers (purchases or returns) each week
"Online" refers to the share of total sales made through the retailer's online store

do not exclude these observations from our analysis, as advertising could influence purchases even for consumers who return more than they purchase during a given time period.[15]

2.3 Sales variance and statistical power

The high variance in the individual data implies surprisingly low power for our statistical tests. Many economists have the intuition that a sample size of a million observations is so large that any economically interesting effect of advertising will be highly statistically significant. Though we had shared this intuition before seeing the results of our experiment, we found that this view 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, we perform an example power calculation. Suppose hypothetically that the campaign were so successful that the firm would have obtained a 100 % short-run return on its investment. The campaign cost approximately R$25,000 to the

[15] In Section 5.3, we will decompose the treatment effect of advertising into its effect on the number of purchasers versus its effect on the purchase amounts conditional on purchase.

retailer,[16] or R$0.02 per treatment-group member, 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 %, a relatively high margin. 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 mean sales per person (R$1.89) during the two-week campaign.

For this hypothetical advertising investment with a 100 % return, can we reject the null hypothesis of no effect of advertising? To answer this question, we note that the standard deviation of two-week sales (R$19) is approximately ten times mean sales and 250 times the hypothesized treatment effect. Even with 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 %. For the alternative hypothesis of a more modestly successful campaign—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 show a surprisingly high probability of type-II error, indicating that the very large scale of our experiment puts us exactly at the measurement frontier where we can hope to detect statistically significant effects of an economically meaningful campaign.[17,18]

3 Treatment effect estimates for the campaign

We estimate the effect of advertising on sales during the two weeks of the campaign by computing the treatment effect on the treated using the exogenous variation from the experiment. Then we show that without an experiment, an observational study of endogenous cross-sectional variation would have yielded spurious results: the point estimate has the opposite sign and three times the magnitude of our experimental point estimate. By contrast, we do not find a large bias with observational panel data, as we next show that a difference-in-differences estimator provides a very similar point estimate. Because the DID estimator also produces smaller standard errors in

[16] Because of the custom targeting to the selected database of known retailer customers, Yahoo! charged the retailer an appropriately higher rate, roughly five times the price of an equivalent untargeted campaign. In our return-on-investment calculations, we use the actual price charged to the retailer for the custom targeting.

[17] This power calculation 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. Note that their sample sizes were closer to 3,000 than to our 1.6 million.

[18] We now realize the importance of doing such power calculations before running an experiment, even one with over a million subjects. Lewis and Rao (2013) detail the statistical imprecision to be expected even in well-designed advertising experiments.

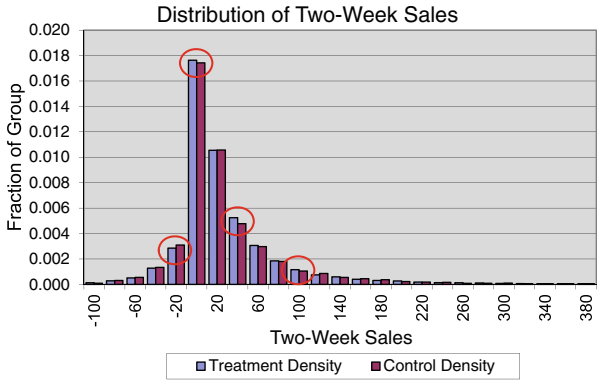

Fig. 3 Sales during the campaign by treatment and control

this setting, we adopt it as our preferred estimate. Finally, we use the experiment to provide tests of the additional identifying assumptions required by DID.

3.1 Estimating the experimental treatment effect on the treated

Although we want to estimate the treatment effect for individuals who are treated with ads, we start with a simple model of the sales difference between the treatment and control groups as a whole, including those individuals who did not browse Yahoo! enough to receive ads:

$$y_i = y_0 + \delta Z_i + \epsilon_i. \tag{1}$$

Here, i indexes the individual, y is sales, Z is the random assignment to the treatment or control group, y_0 is the control group's average sales per person, and ϵ is the error term. Our parameter of interest is δ, the average difference in purchases between the treatment and control groups, which corresponds to the treatment effect on the intent to treat (Angrist et al. 1996).

Figure 3 compares histograms of total sales (online plus offline) during the campaign for the treatment and control groups. For readability, these histograms omit a spike at zero that would represent the 95 % of individuals who made no transaction. We also exclude the most extreme outliers by trimming approximately 0.5 % of the positive purchases from both the left and the right of the graph.[19] Relative to the control, the treatment density has less mass in the negative part of the distribution, corresponding to net returns, and more mass in the positive part of the distribution. Some of the most prominent differences circled in the figure point towards a positive treatment effect.

[19] Out of 75,000 observations with nonzero purchase amounts, we trim about 400 observations from the left and 400 from the right in the histograms. However, we leave all outliers in our analysis, despite their large variance, because customers who account for a large share of sales may also account for a large share of the ad effect. Further, because all data were recorded electronically, we do not suspect typos in the data. Trimming the outliers from the analysis does not appreciably change our results.

Table 4 Mean sales by treatment group and exposure

	Number of observations	Mean sales before campaign (2 weeks)	Mean sales during campaign (2 weeks)	Mean sales difference (During - Before)
Control:	299,426	R$ 1.945	R$ 1.842	-R$ 0.103
		(0.037)	(0.033)	(0.048)
Treatment:	1,277,830	1.934	1.894	-0.039
		(0.018)	(0.017)	(0.024)
Exposed:	814,052	1.813	1.811	-0.002
		(0.021)	(0.021)	(0.029)
Not Exposed:	463,778	2.146	2.042	-0.104
		(0.034)	(0.031)	(0.042)

The descriptive statistics in Table 4 enable us to compute the simple difference between treatment and control groups in Eq. 1. During the campaign the treatment group purchased R$1.894 per person, compared to the control group at R$1.842 per person. The difference of R$0.053 (0.038) per person represents the treatment effect of the intent to treat with ads. The effect is not statistically significant at the 5 % level ($p = 0.162$, two-sided).[20] As a randomization check, we show that before the campaign this difference is much smaller (and opposite in sign): R$1.934 in treatment versus R$1.95 in control, a difference of -R$0.012 (0.041). Both of these sales differences can be viewed as estimates of Eq. 1; we restate them in Table 5 along with the other treatment-effect estimates we are about to compute.

Our simple calculation has estimated the causal effect of the intent to treat with ads (δ), but what we really want to estimate is the treatment effect on the treated. Figure 4 illustrates that 36 % of consumers in the treatment group received no ads due to their endogenous browsing behavior. The randomization ensures that 36 % of the control group also would have received no ads had they been in the treatment group. Ideally, we would remove the unexposed 36 % from both treatment and control groups to eliminate the 36 % dilution and obtain 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.[21] This

[20] However, it does easily exceed the 20 % one-sided significance threshold used to declare a campaign successful in Lodish et al. (1995a).

[21] We recorded zero ad views for every member of the control group and, hence, cannot identify the control group members who would have seen ads. The Yahoo! ad server uses a complicated set of targeting rules and inventory constraints to determine which ad to show to a given individual on a given page. For example, one advertiser's ad might be shown more often on Yahoo! Mail than on Yahoo! Finance. If some other advertiser targeted females under 30 during the same time period, then our ad campaign might have been relatively more likely to be seen by other demographic groups. Our treatment-control assignment represented an additional constraint. We eschew modeling the counterfactual distribution of ad delivery to the control group because we know such modeling would be imperfect and thereby risk biasing our results.

Table 5 Experimental and difference-in-difference estimates

		Difference			Difference-in-Differences						
		(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Model equation: δ or β		1:δ	1:δ	2:β	6:β	6:β	6:β	7:β	8:β	6:β	6,9:β
Treatment-on-treated?		No	Yes	Yes	Yes	Yes	Yes	Yes	Yes/3wk	Yes/8wk	
z[†]	x										
Treatment	Exposed	x[‡]	x	x	x	x	x	x	x	x	x
Treatment	Unexposed	x	x	x	o	–	o	x	o	o	o
Control	Unexposed	o	o	o	o	o	–	o	o	o	o
Dependent variable		y_{pre}	y_{post}	y_{post}	Δy	Δy	Δy	Δy	$\Delta log\ (\bar{y})$[#]	Δy_{3wk}	Δy_{8wk}
Ad effect estimate		-.012	.053	.083	.102**	.101*	.102**	.101	.091**	.166***	.298**
		(.041)	(.038)	(.059)	(.043)	(.055)	(.051)	(.083)	(.041)	(.052)	(.121)

Robust standard errors in parentheses. *, **, and *** represent significance at the 10 %, 5 %, and 1 % levels, respectively
† $N = 1,577,256$ for all models which include all 3 subgroups. In order, $N \approx 814,000$, $N \approx 462,000$, and $N \approx 300,000$ for the three $\{z, x\}$ subgroups
The 8-week estimate includes the follow-up campaign for a total of $N = 868,000$ exposed and $N = 410,000$ unexposed in the treatment group
‡ x = subgroup(s) used to compute lift from ad exposure; o = subgroup(s) used to establish the baseline; - = subgroup(s) omitted from model
# This model comes from equation (8)
Column Notes:
(0) The pre-campaign placebo difference shows no effect
(1) The experimental difference shows an economically meaningful, but statistically weak effect
(2) The local average treatment effect (LATE) estimate simply rescales the experimental difference: $\delta = \frac{\beta}{\hat{\pi}}$ where $\hat{\pi} = 0.64$ (3)
(3) Difference in differences (DID) on the whole sample shows a more statistically significant and slightly larger estimate
(4) DID using the entire control group as a baseline gives a similar, albeit less precise, estimate for the exposed subgroup
(5) DID using the unexposed treatment subgroup as the baseline also gives a similar estimate for the exposed group
(6) LATE estimated on the pre-post differences is equivalent to rescaling the difference between columns (1) and (0)
(7) Using a difference-in-average-log-differences model gives a similar estimate when transformed back to levels
(8) DID using 3 weeks of pre- and post-campaign sales suggests that the effect continues after one week
(9) DID using all 8 weeks of post-campaign data and a rescaling of 3 weeks of pre-campaign data shows continued effects

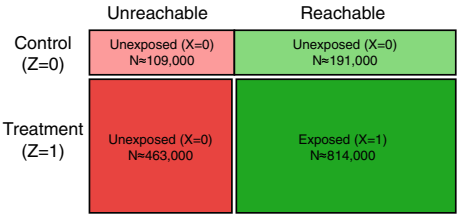

Fig. 4 Experimental design: ad exposure by treatment and control

means we cannot remove the statistical noise of the endogenously untreated individuals; in Section 3.2, we will show that dropping these untreated treatment-group individuals in our cross-sectional comparisons generates considerable selection bias.

However, we can compute an unbiased estimate of the treatment effect on the treated, in spite of the endogenous ad exposure, via a simple rescaling. We define the following system of equations:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{2}$$

$$x_i = \pi Z_i + \nu_i \tag{3}$$

where x_i is an indicator variable for whether individual i saw at least one of the retailer's ads, β is the treatment effect on the treated, π is the fraction of treatment-group users who received at least one ad impression ($= 0.64$), and ν is the first-stage residual. Under mild assumptions β corresponds to the local average treatment effect (LATE; see Angrist et al. 1996) and can be estimated via 2-stage least squares (2SLS) using the randomization, Z, as the instrument. Reducing this system of equations yields a rather intuitive relationship: $\delta = \frac{\beta}{\pi}$. That is, the treatment-on-the-treated estimator is numerically equivalent to merely scaling up our diluted intent-to-treat estimate of R$0.053 by dividing by $\pi = 0.64$, the fraction of individuals treated. This gives an estimate (column 2 of Table 5) of the treatment effect on those treated with ads: R$0.083 (0.059). The standard error also scales proportionally, leaving statistical significance unaffected ($p = 0.162$).

3.2 Spurious results from cross-sectional observational data

Before proceeding, we highlight the kind of spurious result that can occur with non-experimental data. Abraham (2008) reports in the *Harvard Business Review* how comScore uses cross-sectional comparisons of its observations of two million Internet browsers to perform advertising effectiveness studies for clients. 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." In this article, Abraham, like a number of other industry analysts we have met, ignores the endogeneity created by users' media consumption: we should expect correlation of users' browsing behavior with their shopping behavior even in the absence of ads.

If we had estimated the effect of advertising by a cross-sectional observational study in this manner, we would have obtained a very different answer. Without a control group, we would be left with a comparison of endogenously treated versus untreated individuals, equivalent to estimating β (2) without instrumenting for exposure (3). We see from comparing the last two rows of Table 4 that instead of an increase of R$0.083 due to ads, we would 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) would have been reported as highly statistically significant, but we know from the unbiased experimental estimate that this result would have been quite inaccurate.

The large selection bias results from shopping behavior that happens to be correlated with ad views: in this population, those who browse Yahoo! more actively also tend to purchase less at the retailer, independent of ad exposure. We see this clearly in the pre-campaign data in Table 4: 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 that shopping behavior is negatively correlated with Yahoo! browsing and ad delivery, causing a large bias in cross-sectional observational measurement of causal effects—opposite in sign and more than three times larger than the treatment effect.

Knowledge of this selection bias guides our search for a more statistically precise estimator of advertising effectiveness. We would love to exclude unexposed treatment-group members: since advertising has no effect on those who do not see it, these individuals contribute nothing but noise to our estimates. However, simply omitting unexposed treatment-group members would be a big mistake, because the remaining treatment-group members would not represent the same population as the control group. Table 4 shows this selection bias in the pre-campaign data: in the absence of any advertising treatment, to-be-exposed treatment-group members purchase an average of R$1.81, while control-group members purchase an average of R$1.95—a statistically significant difference of -R$0.13 ($p = 0.002$)!

During the campaign, this gap between exposed and unexposed treatment group members becomes considerably smaller than it was in the pre-campaign data. Untreated individuals' sales drop by R$0.10 during the campaign period, while treated individuals' sales remained constant. Control-group mean sales fell by R$0.10 during the same period, just like the untreated portion of the treatment group. The data suggest that advertising had a positive effect that prevented a similar drop in purchases by treated individuals during this time period. This observation leads to our preferred estimator, a difference in differences between treated and untreated individuals, where untreated pools control-group members together with unexposed members of the treatment group.

3.3 Estimating the difference in differences

A difference-in-differences (DID) model leverages the panel nature of our data, using the fact that we observe the same individuals both before and after the start of the ad campaign. We generalize Eq. 2 by indexing time by t and introducing a term α_i to capture the individual heterogeneity in purchases that we have just observed to

be correlated with viewing ads, and a variable τ that represents changes in average sales across time periods. Then we compute a before-after difference to eliminate the individual unobserved heterogeneity α_i:

$$y_{it} = \beta x_{it} + \tau_t + \alpha_i + \epsilon_{it} \quad (4)$$

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta \tau_t + \alpha_i - \alpha_i + \Delta \epsilon_{it} \quad (5)$$

$$\Delta y_i = \beta x_i + \Delta \tau + \Delta \epsilon_i. \quad (6)$$

The difference-in-differences estimator expressed in Eq. 6 involves two time periods: the "pre" period of two weeks before the start of the campaign and the "post" period of two weeks after its start.[22] Since no one saw the ad campaign in the pre-period, $x_{i,pre} = 0$, implying that $\Delta x = x_{i,post}$, which is simply x_i from Eq. 2.

We estimate this difference equation via ordinary least squares (OLS), obtaining a estimate of β directly comparable to the previous treatment-on-the-treated estimate of R$0.083 (0.059). This specification, unlike the specifications in Section 3.1, pools together everyone who saw no ads in the campaign, including both the control group and those treatment-group members who ended up not to seeing any ads. This ends up reducing the standard errors because the split between exposed and unexposed users turns out to be more efficient, close to 50:50 versus the 80:20 split we managed to convince the advertiser to use in the experiment.

Using difference in differences, the estimated average treatment effect of viewing at least one of the retailer's ads during the campaign is R$0.102, with a standard error of R$0.043 (column 3 in Table 5). This effect is statistically significant ($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 $\pm$ 68,000 (95 % confidence interval) due to the campaign. Because the cost of the campaign was approximately R$25,000,[23] the point estimate suggests that the ads produced more than 325 % as much revenue as they cost the retailer. Assuming a 50 % margin on the cost of goods sold, this represents a rate of return of 66 %, a substantial economic impact.

3.4 Evaluating the assumptions of DID

The main identifying assumption of the DID model is that each individual's idiosyncratic tendency α_i to purchase from the retailer is constant over time, and thus the treatment variable is uncorrelated with the error term ($\Delta\tau + \Delta\epsilon_i$) in Eq. 6. 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

[22] Though we have three weeks of pre-period data available, we use only two weeks here for symmetry and simplicity of exposition (two weeks are intuitively comparable to two weeks). Weekly results using a three-week baseline can be found in Table 6.

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

external event, before or during the experiment, had different effects on the retail purchases of those who did versus did not see ads. For example, if the retailer simultaneously launched a direct-mail campaign that was more likely to reach individuals who browsed less often on Yahoo!, such a negative correlation would result in our underestimating the true effect of the ads on Yahoo! The assumption could also be violated if individual-specific shocks (birthdays, vacations, bonuses) were correlated both with purchase behavior and with Yahoo! browsing behavior.[24] Fortunately, our clean experimental estimate of R$0.083 is very similar in magnitude to the DID estimates of R$0.102, reassuring us about the validity of our DID specification.[25] Thus, even when we exploit non-experimental variation in the data, we still make important use of the experiment: the experiment generates more variation in advertising exposure than we would have had in pure observational data and also gives us ground truth against which to check the DID assumptions.

A key difference between our preferred DID estimator and our original treatment-control estimator is that DID compares treated and untreated individuals (pooling the unexposed part of the treatment group with the control group), rather than simply comparing treatment and control groups. We perform a formal specification test of the pooling assumption by comparing pre-post sales differences in the control group versus the unexposed part of the treatment group. The unexposed part of the treatment group has a mean pre-post sales difference just R$0.001 less than that of the control group, and we cannot reject the hypothesis that these two groups' mean differences are the same ($p = 0.988$). As a consequence, DID estimators where the baseline untreated individuals are either the control group or the unexposed part of the treatment group yield nearly identical estimates of R$0.101 (0.056) and R$0.102 (0.051), respectively (columns 4 and 5, Table 5).

We prefer the DID estimate to the simple experimental estimate because it shows a one-third reduction in the standard error of the treatment effect on the treated, from R$0.059 down to R$0.043.[26] Even if the point estimate had remained the same at R$0.083, we still could have claimed statistical significance at conventional levels ($p = 0.054$). But in addition to tighter confidence intervals, we also obtain an increase in the point estimate from R$0.083 to R$0.102. This appears to depend

[24] Lewis et al. (2011) use experiments to show the existence of "activity bias" in observational studies of online-advertising effectiveness: various online activities show high variance across days as well as high correlation across activities, and therefore viewing an ad is positively (but non-causally) correlated with visiting an advertiser's website on a given day. A DID estimator would not correct for such activity bias: with individual-specific shocks, yesterday's incarnation of a person is not a good control for today's. The DID assumption seems a better bet in the present context because we are using lower-frequency data and examining effects on offline (rather than just online) behavior. Further, our specification checks comparing DID to unbiased experimental estimates show that any bias in this setting is considerably smaller than the large effects documented for online behavior by Lewis et al. (2011).

[25] While this specification test is somewhat low-powered, it still tells us that our statistical assumption is plausible. By contrast, we saw above that using endogenous cross-sectional variation fails miserably in a comparison to the clean experimental estimator.

[26] If two estimators are based on valid assumptions, researchers should prefer the estimator with the smallest variance. Selecting the estimator on the variance (second moment) of an adaptive estimator like OLS should not bias the conditional mean.

largely on the (statistically insignificant) differences in sales between treatment and control before the campaign, when the control group purchased slightly more than the treatment group: R$1.945 versus R$1.934. This random one-cent difference in pre-period baseline sales, when properly scaled up using the 0.64 ad exposure rate to R$0.019, accounts for the entire difference between the estimates.

As a robustness check, we compute a difference in differences uncontaminated by non-experimental variation, comparing treatment versus control rather than treated versus untreated users. Formally, we use the the difference equation (6) and first-stage equation (3) from above:

$$\begin{aligned} \Delta y_i &= \beta x_i + \Delta\tau + \Delta\epsilon_i \\ x_i &= \pi Z_i + \nu_i \end{aligned} \tag{7}$$

Here we again use 2SLS. This model (column 6 in Table 5) gives nearly the same point estimate as our preferred DID estimator: intent to treat R$0.064, treatment on the treated R$0.101, confirming that the difference between the experimental and preferred DID estimators is likely due to random differences in pre-period sales between treatment and control. Of course, without the additional efficiency gained from using the non-experimental variation in advertising, this experimental DID estimate remains statistically insignificant at conventional levels ($p = 0.227$). Given the size of the standard errors, we do not wish to emphasize the increase in the point estimate; a conservative reader is welcome to use R$0.083 as their preferred estimate of the treatment effect.[27] We prefer the greater statistical efficiency of the DID model, finding its identification assumptions to be credible.

We note that many types of measurement error could cause our estimates of the effects of advertising to be understated. While we do not know the exact matching algorithm used by the third party, mismatching of sales and advertising data can happen in several ways. For example, the third party who matched the data may have allowed for imperfect matches, such as assuming that two women named Barbara Smith are a single person. Another example is that if a husband browses Yahoo while his wife is logged in to the home computer, we assume she was exposed to the advertising though in fact she was not. In both examples we analyze the customer's sales assuming they were treated, even though we never delivered ads to them. Further, an ad featuring, say, boots at our retailer might stimulate spillover

[27] Though we have modeled the treatment effect as additive, we could instead model it as a constant percentage effect using a difference-in-log-differences estimate at the level of group averages. This could produce a slightly different estimate given that the unexposed group purchases 14 % more than the exposed group, on average, during the baseline pre-period (R$2.06 versus R$1.81). Formally, we write:

$$E[\Delta] = (log(E[y_{E,t}]) - log(E[y_{E,t-1}])) - (log(E[y_{U,t}]) - log(E[y_{U,t-1}])). \tag{8}$$

We estimate $E[\Delta]$=5.0% (2.3%), corresponding to a treatment effect of R$0.091 (0.041) (column 7 in Table 5) computed by multiplying $E[\Delta] \times E[y_{E,t-1}]$. This estimate lies midway between the R$0.083 experimental and the R$0.102 DID estimates.

sales of boots at another retailer, and we would thereby fail to measure the full impact of the advertising. Finally, our results will also underestimate true effects to the extent that our retailer fails to attribute every single purchase to the correct individual customer.

4 Measuring longer-run effects of advertising

We now investigate longer-term effects of the ads after the campaign has ended. One possible case is that the effects could persist, increasing 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 what shoppers would have purchased a few weeks later anyway. Simester et al. (2009), who experimented with a retailer's frequency of catalog mailings, found evidence of such intertemporal substitution, as the firm's best customers responded largely by accelerating purchases forward in time rather than providing truly incremental purchases. In this section, we distinguish empirically between these three competing hypotheses.

4.1 Sales one week after the campaign ended

We first focus on the six weeks of data the retailer provided to analyze the campaign. This includes three weeks of data before, two weeks during, and one week after the campaign. To test 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 following the start of the two-week campaign. For symmetry, we also use all three weeks of pre-period sales, in contrast to the previous section's results using only 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 and untreated treatment-group members).

Using our preferred DID estimator (6), we find that the estimated ad effect increases from R$0.102 (0.043) for two weeks to R$0.166 (0.052) for three weeks (column 8 in Table 5). To isolate the effects in the third week alone, we run DID comparing the third week's sales with the average of the three pre-campaign weeks' sales. This gives 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-week campaign (R$0.051).

4.2 Persistence for more than one week after the campaign

Could the effects persist beyond a week after the campaign ends? We investigate using sales data originally collected to evaluate the follow-up campaign. The retailer provided three weeks of "pre-period data" for the follow-up campaign; these three

weeks fortunately begin immediately after the post-campaign sales data provided for the main campaign.[28]

In order to check for extended persistence of advertising, we use the same DID model (6) as before, estimated on weekly sales.[29] Our "pre-period" sales will be the weekly average of sales in the three weeks preceding the start of the campaign. Our "post-period" sales will be the sales during a given week after the start of the campaign. We then compute a separate DID estimate for each week of the approximately nine weeks after the beginning of the main campaign.[30]

Table 6 displays the results, and Fig. 5 represents them graphically. In the figure, vertical lines indicate the beginning (solid) and end (dashed) of each campaign. The estimated 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$0.035, 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 Fig. 5), we find it striking that every single point estimate is positive.[31] We particularly note the large, positive effects estimated during the inter-campaign period, more than three weeks after ads stopped showing for the campaign. Given the evidence of persistence and the fact that we did not re-randomize between campaigns, we do not attempt to measure separate effects for the follow-up campaign on its own,[32] so we focus instead on the combined effects of both campaigns.

To measure the cumulative treatment effect of both campaigns, we estimate DID on all nine weeks of data following the start of the campaign. We then rescale this number to get a total effect across the entire time period of observation, since the

[28] The campaigns did not start and end on the same day of the week, so we end up with a three-day overlap between the third week after the start of the campaign and the third week prior to the start of the follow-up campaign. That is, those three days of sales are counted twice. We correct for this double-counting by scaling the estimates by the appropriate ratio. In the cumulative 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.

[29] We adapt the model slightly to accommodate varying post-campaign time windows by rescaling the pre-campaign period to be proportional in units of time:

$$\Delta y = y_t - \frac{w_t}{w_{t-1}} y_{t-1} \tag{9}$$

where w_t equals the number of units of time (e.g., weeks) in time window t. For example, if we are comparing a 3-week pre-campaign period to an 8-week post-campaign period, we would use $\Delta y = y_{post}^{8wk} - \frac{8}{3} y_{pre}^{3wk}$. We use this to estimate both the total effects over a multi-week period and separate effects for each week.

[30] Because the follow-up campaign 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 that week 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 is an imperfect, but reasonable approximation, especially considering that the three-day "week" represents such a minor fraction of the long-run period of study.

[31] To avoid overstating the significance of this observation, we note that the weekly estimates are not mutually independent. Each week's DID estimator uses the same three weeks of pre-campaign data, and sales are also modestly correlated from week to week.

[32] Indeed, Table 6 shows that the difference in treatment effect from before to after the start of the follow-up campaign is positive but not statistically significant.

Table 6 Weekly summary of treatment effect on the treated

	Treatment effect*	Robust S.E.
The Campaign		
Week 1 During	R$ 0.047	(0.024)
Week 2 During	R$ 0.053	(0.024)
Week 1 Following	R$ 0.061	(0.024)
Follow-up Campaign		
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 (3 Days)	R$ 0.012	(0.023)
Week 1 Following	R$ 0.004	(0.028)
Average	R$ 0.035	(0.016)

N=1,577,256 customers per week
*When computing the treatment effect on the treated, "treated" individuals are those who have seen at least one ad in either campaign prior to or during that week

"nine-week" time period actually includes a total of only eight weeks.[33] This gives us our eight-week DID estimate of R$0.298 (R$0.121) (column 9 in Table 5).

To estimate the aggregate revenue impact of the campaigns, we multiply our estimate of R$0.298 by the average number of users who had already been treated with ads in a given week. This multiplication by 868,000 gives us a 95 % confidence interval of the total incremental revenues due to ads of R$259,000 $\pm$ 206,000. For comparison, the total cost of these ads was R$33,000. Thus, our point estimate says that the total revenue benefit of the ads was nearly eight times their cost, while even the lower bound of our 95 % confidence interval is two times the cost. Assuming a 50 % margin for the retailer, this lower bound represents break-even. That is, the effect of the two campaigns is statistically significant and economically profitable.

Specification tests for each of the weekly estimates validate our DID estimator. Similar to the specification test computed above for the two-week DID estimate, these tests determine 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 Fig. 6. During each of the 9 weeks following the start of the campaign, the difference in time-series

[33] The first of three weeks prior to the start of the follow-up campaign overlaps with the week following the campaign for three days (see footnote 28). In addition, the follow-up campaign's second "week" is actually only three days, since the campaign ran for only ten days (see footnote 30).

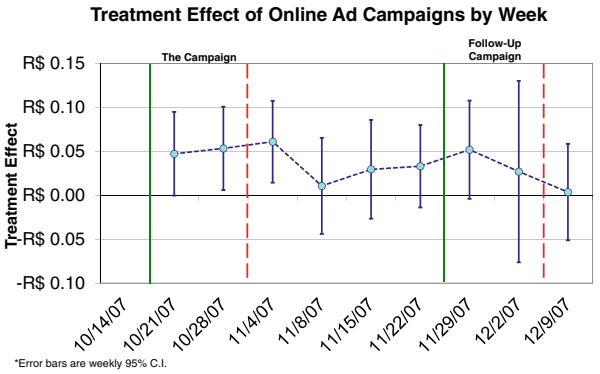

Fig. 5 Weekly DID estimates of the treatment effect

differences between control and untreated treatment group members fails to reject the null hypothesis that the DID model is correctly specified.

4.3 Summary of persistence results

To summarize, 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 accounting for these effects, we find a large return to advertising in our sample period. We still may be underestimating the returns to advertising because our sales data end shortly after the follow-up campaign, so we will miss any additional effects persisting beyond the end of our sample period. The previous reasons for underestimation given in Section 3.4 also remain in force here, so we believe our estimates to be conservative. We hope to further investigate display advertising's persistence in future experiments with longer panels of sales data.

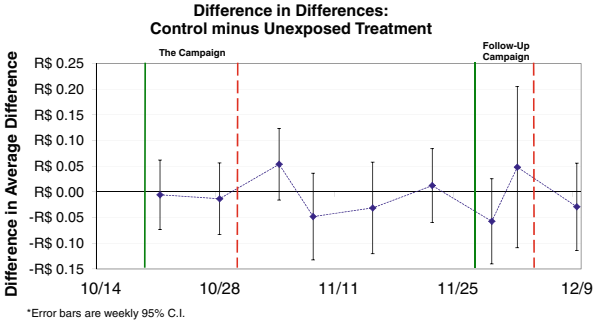

Fig. 6 Weekly DID specification test

5 Decompositions of the treatment effect

In this section, we break down the treatment effects along several dimensions in order to better understand consumer responses to advertising. The questions we address are most intuitively asked about a single campaign; we focus on the first, larger, and more impactful campaign. Despite the evidence on persistent effects in Section 4, we also know longer time differences produce more scope for error in a difference-in-differences specification. To be conservative, we ignore possible sales impacts more than a week after the campaign. In this section, we compute the same difference in differences as in Section 4.1: three weeks before and after the start of the campaign, comparing treated versus untreated individuals.

Our detailed analysis investigates four questions. First, we decompose the effects of online advertising into offline versus online sales, showing that more than 90 % of the impact is offline. Second, we demonstrate that 78 % of the substantial impact on in-store sales occurs for users who merely view but never click the ads. Third, we find that the treatment effect varies with the number of ads viewed by each user. Fourth, we decompose the treatment effect into its impact on the probability of a transaction versus its impact on the average purchase amount conditional on a transaction.

5.1 Offline versus online sales and views versus clicks

In Table 7, we decompose the treatment effect into offline and online components by estimating DID separately for offline and online sales. The first line shows that most of the treatment effect comes from brick-and-mortar sales. The treatment effect of R$0.166 per treated individual consists of a R$0.155 effect on offline sales plus a R$0.011 effect on online sales. That is, 93 % of the online ads' total effect was on offline sales. This result will surprise those who assume that online ads only impact online sales. In fact, the effect is approximately proportional to the relative size of the two sales channels; the data in Table 3 indicate that approximately 85 % of all sales are offline.

In online advertising, the click-through rate (CTR) quickly became a standard measure of performance, automatically providing much more information than is

Table 7 Offline/online and viewer/clicker ad effect decomposition

	Total sales	Offline sales	Online sales
Ads viewed (β, Eq. 6)	**R$ 0.166**	**R$ 0.155**	R$ 0.011
[63.7 % of Treatment group]	(0.052)	(0.049)	(0.016)
Ads viewed, not clicked (β_0, Eq. 10)	**R$ 0.139**	**R$ 0.150**	-R$ 0.010
[92.8 % of Viewers]	(0.053)	(0.050)	(0.016)
Ads clicked (β_1, Eq. 10)	**R$ 0.508**	R$ 0.215	**R$ 0.292**
[7.2 % of Viewers]	(0.164)	(0.157)	(0.044)

DID estimates; bold denotes statistical significance at the $\alpha = 0.05$ level

available in traditional media campaigns. Still, the CTR does not measure what advertisers care most about: the impact on sales. Furthermore, average click-through rates have fallen by an order of magnitude during the short history of online display advertising, from 1.1 % in 1998 to 0.09 % in 2008.[34] An interesting question is, therefore, "To what extent do ad clicks capture the effects of advertising on retail sales?"

We answer this question in the second and third lines in Table 7 by adding heterogeneous treatment effects to our DID estimator:

$$\Delta y_i = \beta_0 x_i \cdot 1(Clicks = 0) + \beta_1 x_i \cdot 1(Clicks > 0) + \Delta\tau + \Delta\epsilon_i. \quad (10)$$

We partition the set of treated individuals into those who clicked on an ad (β_1, line 3) versus those who merely viewed ads but did not click any of them (β_0, line 2). 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 treatment effect estimate 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 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. By contrast, for online sales, those who view but do not click have a treatment effect precisely measured to be near zero, so we can conclude that ads do not cause non-clickers to buy much more online. Turning to the set of clickers, we see that ads cause a large difference in purchase amounts 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.

Finally, we note that DID is what makes our click-versus-view decomposition possible. The experiment generated exogenous variation in views but not in clicks. Clicks are fundamentally endogenous behavior, so comparing the levels of sales for clickers versus non-clickers would have been biased by selection effects: those shoppers intrinsically likely to purchase the most are probably also the most likely to click an ad. We assume that this heterogeneity is constant over time in order to use before-after differences to compare the impact of advertising on clickers versus non-clicking viewers.

[34] See Meland (1999), Holahan and Hof (2007), and Shein (2012) for data on the historical decline in CTR. At 0.3 %, our highly targeted advertising campaign had a rather high CTR, three times that of the average display campaign in 2007.

5.2 Number of ad exposures

Recalling from Fig. 2 that ad exposure widely varies across individuals, we now ask how the treatment effect varies with the number of ads seen. We nonparametrically regress the 3-week pre-post difference in sales on the number of ad views during the campaign, estimating a variant of equation (6):

$$\Delta y_i = \beta(f_i) + \Delta\tau + \Delta\epsilon_i, \tag{11}$$

where $\beta(\cdot)$ is an unknown advertising response function and f_i is the number of ad exposures for individual i. We use locally quadratic regression with the Epanechnikov kernel and a bandwidth of 15 ad views to estimate $\beta(\cdot)$. The treatment effect iszero for those who did not view ads, so we normalize the curve's intercept to equal zero for those with zero ad views.[35] Figure 7 shows the nonparametric estimate with 95 % pointwise confidence intervals.

We see that the treatment effect is initially strongly increasing in the number of ad views. A parametric linear regression on the range from 0 to 50 ad views gives a slope of R$0.0099 (0.0022) per impression. The effect peaks just under R$0.40 at 50 ads and hovers near this level until 100 impressions per person. Beyond this, the data becomes so sparse (only 6.1 % of the treatment group receives more than 100 impressions) that the effect is no longer statistically distinguishable from zero.

We caution that this graph may not represent the causal effect of increasing the number of ads shown to a given individual. This causal interpretation could easily be invalid because the number of ad views was not exogenously varied by individual. Each individual has a browsing behavior "type" that determines the distribution of pages they visit on Yahoo! and, indirectly, the number of ads that user receives. We know from the previous results in Table 4 that browsing behavior on Yahoo! is (negatively) correlated with retail purchases in the absence of advertising, so we shy away from a causal interpretation. We are on solid ground only when we interpret the graph as displaying heterogeneous treatment effects by the user's browsing type.

How do the incremental revenues compare with the cost of delivering ads to each type of user? The upward-sloping line on the graph represents the retailer's per-person cost of purchasing a given number of ad impressions. This line has a slope of R$0.001, the retailer's price per impression. Thus, the graph shows 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 impressions is a break-even 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 ads.

So, what are the incremental profits? To transform the incremental-revenue curve into profits, we again assume a 50 % retail profit margin and multiply the entire incremental-revenue curve by 50 %, reducing its vertical height by half. Due to this empirical curve's shape, the break-even point turns out to be nearly the same: around 100 ads per person. The retailer might gain from a policy that caps the number of ad

[35] Because we cannot observe counterfactual ad views for the control group, we must rely on DID, pooling control-group members with untreated treatment-group members.

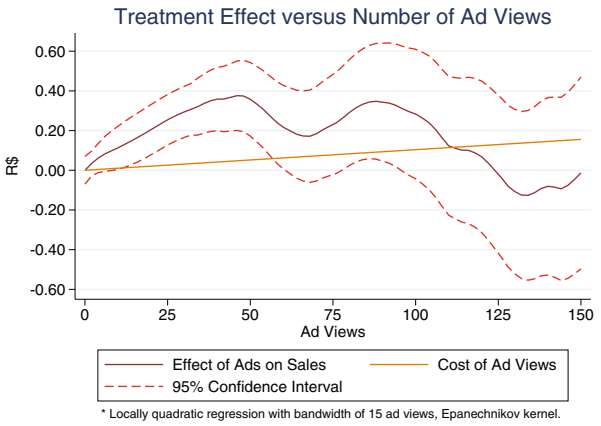

Fig. 7 Nonparametric estimate of the treatment effect by ad views

views per person at 100 by avoiding spending money on individuals for whom the benefits may be less than the cost of the ads. This hypothesis could fruitfully be investigated in future experiments, as could hypotheses about other types of heterogeneous treatment effects. In a companion paper which only uses the experimental variation (Lewis and Reiley 2014), we show that 38 % of the treatment effect from this advertising campaign comes from the oldest 6 % of consumers. Using DID to estimate heterogeneous treatment effects for age and gender (whose details we omit from this paper for the sake of brevity), we find similar but statistically stronger results. Thus, an advertiser could increase effectiveness per delivered online ad by using experiments to identify better targeting strategies (see also Lewis and Reiley 2012).

5.3 Probability of purchase versus basket size

We decompose advertising's average effect on purchases into two separate components of interest to retailers: the effect on the probability of a transaction versus the effect on the "basket size," or purchase amount conditional on a transaction. 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 among those who purchased. The product of these two numbers gives the unconditional average purchase amount of R$2.64 per person. We reproduce these numbers in Table 8 alongside the treatment effect estimates from a three-week pre-post DID model for each variable of interest.

First we investigate advertising's impact on the probability of a transaction.[36,37] The first row of Table 8 shows the statistically significant ($p = 0.03$) increase of

[36] 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, this effect would likely be larger if we restricted our analysis to positive purchases.

[37] We present a simple DID in sample proportions: our results are comparable to an OLS linear probability model rather than to a nonlinear model like a probit.

Table 8 Basket size and frequency decomposition of ad effect

	3-Week DID treatment effect	Treated group level*
$Pr(Transaction)$	0.10 %	6.48 %
	(0.05 %)	
Mean basket size	R$ 1.75	R$ 40.72
	(0.74)	
Revenue per person	R$ 0.166	R$ 2.639
	(0.052)	

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

0.102 % in the probability of purchase due to the advertising. This represents an increase of 1.6 % in the average probability of a transaction.

Next we consider the effect on basket size. Since the sales data are sparse and most purchasers do not purchase in both time periods, we cannot employ the same customer-level DID estimator as before. Instead, we compute DID using group means (for each group, the mean over all nonzero purchase amounts) and pay careful attention to possible time-series correlation when computing standard errors.[38] As shown in the second row of Table 8, the ad 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 R$40.72, this represents an increase of 4.5 %.[39]

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 and an increase of R$1.75 in the purchase amount conditional on a transaction, representing percentage increases over 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. We hope these empirical observations will eventually contribute to a deeper understanding of the mechanisms by which advertising affects consumers.

6 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 make progress in this measurement problem by conducting a large-scale field experiment

[38] When comparing mean time-series differences between treated individuals and untreated individuals, those two means are independent, so standard errors are straightforward. But when computing DID for four group means, pre- and post-campaign basket-size estimates are correlated from some customers purchasing in both periods.

[39] Because the advertising increases the number of purchasers, the change in average basket size conflates two effects: the change in inframarginal customers' purchase amounts and any difference in marginal and inframarginal customers' average purchase amounts.

that systematically varies advertising to over one million retail customers on Yahoo! Even with such a large individual-level dataset, we have just barely reached the frontier of measuring 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 be detected with only a 63 % probability. Sales at this retailer have high variance, and this online advertising campaign is just one of many factors that influence purchases. These facts make the treatment-control differences noisy. For more precise estimates, we compute a difference in differences using a panel of weekly individual transactions, exploiting both experimental and non-experimental variation in ad exposure. This DID estimator requires more assumptions than the simple experimental difference, but both estimators give similar point estimates.

Our primary result is that this advertising was profitable for the retailer. We find positive, sizable, and persistent effects of online retail advertising on sales. The effects appear to persist for several weeks after the last ad was shown. In total, we estimate that the retailer's incremental revenues were more than seven times the cost of the ads.

Though some people assume that online advertising mostly impacts online retail sales, we find the reverse to be true. This retailer records 85 % of its sales volume offline, and we estimate 93 % of the treatment effect is on offline sales. Online advertising evidently can have a large effect on offline sales.

Even though clicks are a standard measure of performance in online-advertising campaigns, we find that focusing only on clickers leads to a serious underestimate of the campaign's effects. 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 ad effect is largest for users who saw between 25 and 100 ad impressions during the campaign. We also find that online advertising increases both the probability of purchase and the average purchase amount, with three-quarters of the treatment effect coming through increases in the average purchase amount.

One of our most important results is a demonstration of just how poorly one can measure the causal effects of advertising using cross-sectional variation. If we had neither an experiment nor panel data, but instead attempted to estimate these effects using cross-sectional variation in endogenous ad exposure, we would have obtained a result opposite in sign to the true estimate. Futhermore, the magnitude of the selection bias would be more than three times the size of the true measured effect of advertising.

In this experiment, we find a negative correlation between ad exposure and baseline purchasing, but we believe that most cross-sectional observational studies are likely to find a positive correlation, thereby leading to overestimated effects of advertising (see, for example, the comScore methodology described in Abraham (2008)). Our sample includes only those users whom the advertiser intended to target with the advertising campaign. In most observational studies the researcher will not observe exactly which users were targeted. Because advertisers usually target ads to those

most likely to be interested in the product, we should expect higher baseline purchases from exposed than unexposed users. 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, we can easily observe a positive, non-causal correlation between advertising and consumer demand for the advertised good.

After multiple years of interactions with advertisers and advertising sales representatives over such measurement issues, we have come to believe that many advertisers do not have credible estimates of the effects of their brand advertising (for example, see eBay's recent experience in Blake et al. (2013)). Most advertisers do not systematically vary their levels of advertising to measure its causal effects, nor do they have good individual-level panel data to account for unobserved heterogeneity. Thus, most practitioners are unable to overcome problems of endogeneity, selection, and omitted-variable bias, nor do they even have a strong sense of the likely magnitude of the biases. We have demonstrated that one technique used by practitioners can easily have a bias three times the size of the effect one is attempting to estimate.

Another important result is our validation of the use of panel data in observational studies of advertising effectiveness. Implementing difference in differences on the panel data gives us a very similar estimate to the experimental difference, and the DID estimate is robust to using different subsets of the unexposed users as our pseudo-control group. 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. The close match between our experimental estimator and our DID estimator is good news for analysts who are unable to generate experiments but have panel data on a large sample of exposed versus unexposed users. However, we caution that this close match may easily fail in other settings. In particular, the problem of online activity bias pointed out by Lewis, Rao and Reiley (2011; see footnote 24 above for details) suggests that for many online outcomes a panel estimator would fail to eliminate bias. We also note that experiments can augment statistical power in an observational study: in this case, the experiment created 200,000 of the 700,000 unexposed users.

In future research, we hope to replicate these results with other retailers. We are using what we have learned in this study to design better experiments. For example, future experiments should carefully mark control-group members who could never have been exposed to the ads; excluding these observations will give more efficient estimates of treatment effect on the treated. We know that occasional attribution of purchases to the wrong customer almost certainly biases our estimates towards zero, and we hope to reduce (or at least assess the size of) this bias through improved measurement. We also hope 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.

Acknowledgments We thank Meredith Gordon, Sergiy Matusevych, and especially Taylor Schreiner for their work on the experiment and data. Yahoo! Inc. provided financial and data assistance and guaranteed 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, David Broockman, JP Dubé, Liran Einav, Glenn Ellison, Matt Gentzkow, Jerry Hausman, Kei Hirano, Garrett Johnson, Larry Katz, John List, Preston McAfee, Sendhil Mullainathan, Justin Rao, Paul Ruud, Michael Schwarz, Pai-Ling Yin, and many others, including attendees at many conferences and seminars.

References

Aaker, D.A., & Carman, J.M. (1982). Are you overadvertising? a review of advertising-sales studies. *Journal of Advertising Research*, *22*(4), 57–70.

Abraham, M.M. (2008). The off-line impact of online ads. *Harvard Business Review*, *86*(4), 28.

Abraham, M.M., & Lodish, L. (1990). Getting the most out of advertising and promotion. *Harvard Business Review*, *68*(3), 50.

Ackerberg, D. (2003). Advertising, learning, and consumer choice in experience good markets: an empirical examination*. *International Economic Review*, *44*(3), 1007–1040.

Ackerberg, D.A. (2001). Empirically distinguishing informative and prestige effects of advertising. *RAND Journal of Economics*, 316–333.

Ackoff, R.L., & Emshoff, J.R. (1975). Advertising research at anheuser-busch, inc. (1963-68). *Sloan Management Review (pre-1986)*, *16*(2), 1–1. http://search.proquest.com/docview/206793115?accountid=12861.

Allaire, Y. (1975). A multivariate puzzle: a comment on advertising research at anheuser-busch, inc. (1963-68). *Sloan Management Review, (Spring)*, *91*, 94.

Angrist, J.D., Imbens, G.W., Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444–455.

Bagwell, K. (2008). The economic analysis of advertising. *Handbook of industrial organization*, *3*, 1701–1844.

Berndt, E.R. (1991). *The practice of econometrics: classic and contemporary*. Reading: Addison-Wesley.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., Zinman, J. (2010). What's advertising content worth? evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, *125*(1), 263–306.

Blake, T., Nosko, C., Tadelis, S. (2013). Consumer heterogeneity and paid search effectiveness: a large scale field experiment. *NBER Working Paper*, 1–26.

DellaVigna, S., & Gentzkow, M. (2010). Persuasion: empirical evidence. *Annual Review of Economics*, *2*(1), 643–669.

Eastlack, J., & Rao, A. (1989). Advertising experiments at the campbell soup company. *Marketing Science*, 57–71.

Ghose, A., & Yang, S. (2009). An empirical analysis of search engine advertising: sponsored search in electronic markets. *Management Science*, *55*(10), 1605–1622.

Holahan C., & Hof R.D. (2007). So many ads, so few clicks, Bloomberg Businessweek. http://www.businessweek.com/stories/2007-11-11/so-many-ads-so-few-clicks.

Hu, Y., Lodish, L.M., Krieger, A.M. (2007). An analysis of real world tv advertising tests: a 15-year update. *Journal of Advertising Research*, *47*(3), 341.

Levitt, S.D., & List, J.A. (2009). Field experiments in economics: the past, the present, and the future. *European Economic Review*, *53*(1), 1–18.

Lewis R.A., & Rao J.M. (2013). On the near impossibility of measuring the returns to advertising. Working paper.

Lewis, R.A., & Reiley D.H. (2014). Advertising effectively influences older users: how field experiments can improve measurement and targeting. *Review of Industrial Organization* forthcoming. http://rd.springer.com/article/10.1007/s11151-013-9403-y.

Lewis, R.A., Rao, J.M., Reiley, D.H. (2011). Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web* (pp. 157-166).

Lewis, R.A., Reiley, D.H., Schreiner T.A. (2012). Ad attributes and attribution: large-scale field experiments measure online customer acquisition. Working Paper.

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

Lodish, L.M., Abraham, M.M., Livelsberger, J., Lubetkin, B., Richardson, B., Stevens, M.E. (1995b). A summary of fifty-five in-market experimental estimates of the long-term effect of tv advertising. *Marketing Science*, *14*(3 supplement), G133—G140.

Meland, M. (1999). Banner click-throughs continue to fall. Forbes. http://www.forbes.com/1999/05/11/mu8.html.

Shein, E. (2012). Banner ads: past, present, and... future? CMOcom. http://www.cmo.com/content/cmo-com/home/articles/2012/4/24/banner-ads-past-present-and--future.html.

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