Measuring the Effects of Online Advertising on Human Behavior Using Natural and Field Experiments

by

Randall A. Lewis

B.A. Economics, B.S. Mathematics
Brigham Young University, 2006

SUBMITTED TO THE DEPARTMENT OF ECONOMICS IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN ECONOMICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2010

Signature of Author: ______________________________
Department of Economics
August 15, 2010

Certified by: ______________________________
Glenn Ellison
Gregory K. Palm (1970) Professor of Economics
Thesis Supervisor

Certified by: ______________________________
Jerry Hausman
John & Jennie S. MacDonald Professor of Economics
Thesis Supervisor

Accepted by: ______________________________
Esther Duflo
Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics
Chairman, Departmental Committee on Graduate Theses

Measuring the Effects of Online Advertising on Human Behavior Using Natural and Field Experiments

by

Randall A. Lewis

Submitted to the Department of Economics on August 15, 2010
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Economics

ABSTRACT

This thesis investigates the effects of online advertising on human behavior: clicks, new-account sign-ups, and retail sales. Five chapters cover natural and field experiments used to measure these effects for both display and search advertising.

The first chapter uses a natural experiment on the Yahoo! Front Page, aided by a flexible semiparametric model, to identify the *causal* effects of display ad frequency on internet users' responses as measured at the individual level by clicks and new-account sign-ups. Performance is heterogeneous regarding frequency and clickability; some campaigns exhibit significant decreasing returns to scale after one or two impressions while others show constant returns to scale even after fifty impressions. For some campaigns, a simple nonparametric regression which ignores selection bias finds increasing returns to scale, but none is found with the model that uses exogenous variation in views. Conversely, many campaigns that appear to exhibit diminishing returns when failing to account for selection, in fact, show little to no wear-out.

The second chapter assesses the ability of online display advertising to attract new customers by analyzing a large-scale field experiment which exposed 3.7 million subjects to ads on Yahoo!. The number of new account sign-ups at an online business was tracked and shows a statistically significant impact of one of the two types of advertising campaigns. The ads served as Yahoo! run-of-network succeeded in generating a statistically significant increase in sign-ups of 8-14% relative to the control group. The ads shown on Yahoo! Mail did not produce a statistically significant increase in sign-ups. Despite being derived using millions of subjects, this estimate is quite noisy, with the upper bound of the 95% confidence interval estimate being a 15% increase in new customers. These estimates call into question click-only attribution models, as the number of users that clicked on an ad and converted is less than 30% of the estimated treatment effect.

The third chapter asks, "Does advertising affect sales in a measurable way?" New technologies for tracking both sales and advertising at the individual level are used to investigate the effectiveness of brand advertising for a nationwide retailer. A controlled experiment on 1,577,256 existing customers measures the *causal* effect of advertising on actual purchases, overcoming the major hurdles regarding attribution typically encountered in advertising effectiveness research by exogenously varying exposure to the ads. Each customer was randomly assigned to treatment and control groups for an online advertising campaign for this retailer. Online brand advertising generated a statistically and economically significant effect on in-store sales for this retailer. The design of the experiment permits a demographic breakdown of the advertising's heterogeneous effects. Somewhat surprisingly, the effects are especially large for the elderly. Customers aged 65 and older, comprising only 5% of the experimental subjects, exhibited a 20% average increase in sales due to the advertising campaign, which represents 40% of the total effect among all age groups.

The fourth chapter further investigates the effects of online advertising on sales. A quasi-experimental approach is taken to analyze the randomized experiment in Chapter 3. Individual-level data on ad exposure and weekly purchases at this retailer, both online and in stores, are combined and used to find statistically and economically significant impacts of the advertising on sales. The treatment effect persists for weeks after the end of an advertising campaign, and the total effect on revenues is estimated to be more than seven times the retailer's expenditure on advertising during the study. Additional results explore differences in the number of advertising impressions delivered to each individual, online and offline sales, and the effects of advertising on those who click the ads versus those who merely view them.

The fifth chapter quantifies the externalities exerted among competing north ads in search advertising. "North" ads, or sponsored listings appearing just above the organic search results, generate the majority of clicks and revenues for search engines. The research question asks, "Does increasing the number of rival north ads decrease the number of clicks I receive on my own north ad?" A controlled experiment investigates this question and finds, surprisingly, that additional rival ads in the north tend to increase rather than decrease the click-through rate (CTR) of the top sponsored listing. Several possible explanations are proposed for this phenomenon, and directions for new theoretical models of sponsored search are suggested.

Thesis Supervisor: Glenn Ellison
Title: Gregory K. Palm (1970) Professor of Economics

Thesis Supervisor: Jerry Hausman
Title: John & Jennie S. MacDonald Professor of Economics

Acknowledgements

I would like to express my gratitude for the many people who have supported my research at the MIT Economics Department. I would especially like to thank Glenn Ellison for his encouragement as he helped me see how to improve my writing. His instruction, both in classes and for research, has been invaluable. I would like to thank Jerry Hausman for his support of my econometric and advertising research over the past four years. Stephen Ryan and Whitney Newey provided valuable support to direct my research. Josh Angrist, Sara Ellison, Panle Jia, Nancy Rose, Richard Schmalensee, Catherine Tucker, and other attendees of the MIT IO lunches provided valuable feedback, even during my time spent in California.

I also thank my MIT classmates. They have challenged and supported my research since our very first semester. Michael Powell, Tatyana Deryugina, Mar Reguant-Rido, Daniel Keniston, and Brigham Frandsen each contributed in unique and numerous ways.

Yahoo! Research has been both flexible and strongly supportive over the past two years. I specifically thank Michael Schwarz, Preston McAfee, and Justin Rao for their encouragement and feedback on each of these chapters.

This acknowledgement would be immensely incomplete if I failed to thank David Reiley. He has been my research advocate during the past two years—patient throughout my mistakes and misunderstandings while championing and publicizing our triumphs. Thank you, David!

Lastly, I would like to thank my family. My parents have always been supportive of my academic pursuits and been actively involved in facilitating my educational opportunities. My wife, Clara, provided countless hours of support, patience, and conversation. She was willing to “ride the emotional rollercoaster” of research—both successes and failures. She was willing to take the plunge and experience an adventure in Cambridge while I studied at MIT and was patient throughout my frustrations, challenges, and successes there. She was willing to take a risk and move to California so that I could intern at Yahoo!. She was willing to leave her friends behind during my fourth year so that I could focus on this research while living in Silicon Valley. And, through it all, she has never complained about these hardships but has been a true friend. For that, I cannot sufficiently express my appreciation. Thank you!

Table of Contents

Chapter 1

Where's the "Wear-Out?"

Online Display Ads and the Impact of Frequency

Randall A. Lewis*

Abstract

This paper uses a natural experiment on the Yahoo! Front Page, aided by a flexible semiparametric model, to identify the *causal* effects of frequency on internet users' responses as measured at the individual level by clicks and new-account sign-ups. Performance is heterogeneous regarding frequency and clickability; some campaigns exhibit significant decreasing returns to scale after one or two impressions while others show constant returns to scale even after fifty impressions. For some campaigns, a simple nonparametric regression which ignores selection bias finds increasing returns to scale, but none is found with the model that uses exogenous variation in views. Conversely, many campaigns that appear to exhibit diminishing returns when failing to account for selection, in fact, show little to no wear-out.

JEL Classification

D83 - Information and Knowledge, L86 - Information and Internet Services, M37 - Advertising

Keywords

Online display advertisement, frequency, reach, identification, advertising effectiveness, natural experiment, false synergy, wear-out

* Lewis: Massachusetts Institute of Technology and Yahoo! Research, randallL@mit.edu and ralewis@yahoo-inc.com. I thank Daniel Suh from the Yahoo! Front Page team for answering questions and providing information which made this research possible. Yahoo! Inc. provided financial and data assistance while guaranteeing academic independence prior to the analysis. I appreciate helpful comments from my colleagues Justin Rao, David Reiley, and Michael Schwarz at Yahoo! Research as well as Clara Lewis, Glenn Ellison, Jerry Hausman, Stephen Ryan, Nancy Rose, and participants in the MIT Industrial Organization Lunch Workshop. Thank you!

I. Introduction

"Repetition is an important part of advertising. Repetition is an important part of advertising."

BusinessKnowledgeSource.com [1]

"The value of an ad is in inverse ratio to the number of times it has been used."

Raymond Rubicam, American Advertising Pioneer

Online advertising is all about repetition. Tens of billions of dollars are spent each year on repetitive online media such as email, search, display, radio, and video ads intended to reach consumers with information and enticements. In 2009, $22.7 billion were spent advertising online (IAB Internet Advertising Revenue Report 2009) with display ads accounting for 22% or $5.1 billion. Despite the scale of these advertising expenditures and the voluminous repetition of identical messages to consumers, little is known about the marginal effectiveness of these messages. Advertisers need to understand how exposure frequency influences consumer behavior. Frequency forms the primary determinant of the cost of advertising: how many times can I send the message and still obtain a profitable return? As the number of messages increases, the advertiser can experience increasing, decreasing, or constant returns to scale. Increasing returns to scale is commonly referred to as "wear-in" or "synergy" while decreasing returns is known as "wear-out" (Pechmann and Stewart 1990).

The main challenge facing a study of ad frequency is that advertisers intentionally target their ads to responsive subpopulations. As a result, the people that are thought to be more likely to respond to the ad are more likely to see a large number of impressions. This confounds the analysis of frequency and tends to bias estimates toward synergy. Past studies have attempted to overcome this problem by using a controlled laboratory setting or trying to account for selection using econometric procedures. While this research has furthered our understanding of advertising, both groups have well-known drawbacks. Laboratory studies have been criticized due the problem of external validity. Observational studies have a difficult time effectively controlling selection bias. In this paper, I use a natural experiment described in section II that induced exogenous variation in ad frequency while retaining the natural decision-making environment of the field. Visitors to the Yahoo! Front Page are shown a different ad depending

[1] http://www.businessknowledgesource.com/marketing/the_importance_of_repetition_in_advertising_021569.html

on whether they arrive on an even or odd second. This pseudo-random ad exposure, when coupled with data on clicks and sign-ups for the same users, allows me to identify the causal effects of frequency on these outcomes for a number of online display ad campaigns.

The Yahoo! Front Page is a major internet portal where 30-40 million users visit each day. On every visit, the primary ad unit is refreshed and the user is delivered another impression. The even and odd second rotation of two advertisers on certain days induces exogenous variation in the number of ads each user sees for each of the two advertisers. However, the total number of ads seen is not exogenous—it is determined by the number of times the user decides to visit www.yahoo.com. If the number of visits to the Yahoo! Front Page is correlated with unobserved heterogeneity that influences responsiveness to ads, then models which fail to account for this heterogeneity will, in general, produce biased estimates of the impact of frequency.

I have individual-level data on ad views and ad clicks for 30 campaigns shown on 15 days from January through March 2010 on the Yahoo! Front Page. I develop and present a model in section III to estimate the effects of frequency on clicking behavior. The model controls for the unobserved heterogeneity in number of visits when estimating the effects of frequency on the likelihood of a user to click on the ad to follow the link to the advertiser's website. As a result, the model only exploits the exogenous variation in the number of ads seen by users who visit the page a given number of times.

While clicks are commonly analyzed by researchers, conversions are more interesting to study because they represent economic activity which directly affects the firm's bottom line. To accommodate conversions, I adjust the restrictions on the model to estimate the effects of frequency on an individual's likelihood of signing up for a new account from being exposed to that campaign. The online environment allows me to track both the number of impressions seen as well as whether the user triggered a beacon[2] on the advertiser's website. I combine four of the advertiser's campaigns to examine the effects of frequency on conversions.

I find in sections V and VI that there is widely-varying heterogeneity in the effects of frequency on clicker rates. Four of the 30 campaigns wear out exceedingly fast, while ten exhibit near constant returns to scale for more than 20 impressions. For example, several campaigns achieve roughly 40 times the impact of showing one ad from showing 40 ads while other

[2] A beacon is also known as a pixel because, in practice, it is a 1x1 pixel graphic downloaded by the user used to record the visit by that user's browser cookie to a website.

campaigns achieve only 3 times the impact of showing one ad from showing 40. I show that naively using observational data and ignoring selection bias can lead to large biases in the frequency estimates; for clicking on the Yahoo! Front Page, the bias overstates wear-out for 26 of the 30 campaigns. In addition to sizeable over- and under-statements of the estimates of wear-out, erroneous findings of synergy (increasing returns to scale) are also common—half of the campaigns find synergy from showing two ads when compared to showing just one. Using exogenous variation in ad delivery and controlling for user type, I find no evidence of synergy. The analysis of frequency on conversions shows a positive return to frequency for as many as 20 impressions.

Past literature has made valuable progress understanding the effects of ad frequency using a variety of empirical techniques. An early influential field experiment by Zielske (1959) measured the impact of direct mail ad frequency on housewives' recall of a low-market-share brand of flour over the course of a year. Each woman was mailed thirteen ads, each spaced by either one week or four weeks. Brand-name recall decayed exponentially over time, and four-week intervals yielded a higher average brand-awareness over the course of the year than did the one-week intervals—a result consistent with decreasing returns to frequency on recall (Simon, 1979). Lab experiments have also been used to study the effects of repetition of print media on recall (Craig, Sternthal, and Leavitt, 1976) and of television and internet ads on attitudinal measures (Campbell and Keller, 2003). Their findings corroborate Zielske's field experiment in finding decreasing returns to frequency—and potentially even negative marginal effects of frequency from high levels of exposure.

In a pioneering field study with observational data, Tellis (1988) combined data on television ad exposure with scanner data for consumer purchases to evaluate the impact of advertising frequency on sales. Using a Tobit analysis to correct for selection biases, he concludes that the optimal frequency for television is perhaps two or three commercials per week. However, he cautions that extending the results beyond those frequencies is difficult because the study lacks sufficient data to make any claims beyond four or five exposures.

With the widespread adoption of the internet in the mid 1990s and the proliferation of online display and search advertising to support this new service medium, many researchers harnessed new technology to more effectively explore the impact of ad frequency on behaviors. Early observational research by Chatterjee, Hoffman, and Novak (2003) studied click-through rates (CTRs), or the fraction of ads downloaded that were clicked on by users in order to visit the advertiser's website. They found wear-out in that increasing frequency led to a decline in CTRs. Kameya and Zmija (2002) outlined other early research using observational data which found that after five impressions, returns are diminishing (Carlon & Hislop, 2001; Morgan Stanley Dean Witter, 2001). Dreze and Hussherr (2003) used surveys and eye-tracking equipment to study online display advertising in 1999. They found significant "ad blindness" among participants, where users had learned to avoid looking at certain areas of the webpage because they knew what looked like an ad versus the content that they were seeking. In particular, they found that the observed 50% likelihood of users looking at the ad was sufficient to influence consumers' brand awareness and recall. Havlena, Cardarelli, and de Montigny (2007) combined online display advertising frequency measurements with TV and print advertisement frequencies and found positive returns to frequency on several survey-based measurements such as recall.

More recently, using a field experiment targeted at a matched sample of Yahoo! visitors and a nationwide retailer's customers, Lewis and Reiley (2010a) found that online display advertising produced a statistically and economically significant positive effect on online and offline sales. Yet, those who visited Yahoo! and saw the retailer's ads actually purchased less, on average, than those who were eligible to see the ads but failed to browse online enough to see any. Not surprisingly, then, a naïve regression of each customer's sales on the number of ads seen produced a negative slope, falsely suggesting that advertising stifles sales. This analysis, similar to many recent field studies, incorrectly assumes that the number of ads seen is exogenous even though users who visit Yahoo! more frequently are shown more ads. Ongoing research by Lewis and Reiley (2010b) measures the causal impact of frequency on both online and offline sales by building upon the work in this paper.

In summary, research has improved our understanding about the impact of advertising in a variety of media; however, several shortcomings of lab experimentation and empirical studies using observational data have limited the solidarity of the conclusions regarding the effects of frequency for display advertising. Field experiments, whether natural or not, are the best way to cleanly study the effects of frequency. By identifying exogenous variation in the number of ads shown to each person and observing their behavioral responses, I can accurately identify the effects of frequency.

II. Yahoo! Front Page Ads & Identification Strategy

Identification of the causal effects of advertising requires exogenous variation in the number of ad impressions. The main rectangular ad unit on the Yahoo! Front Page provides a unique natural experiment where the number of ads shown to users is random, given their endogenous number of visits, and responses to the ads such as clicks and sign-ups can be tracked at the individual level. Here I outline this source of the exogenous variation and provide relevant details about the natural experiment.

The Yahoo! Front Page for the United States market, shown in Figure 1-1, is a major internet portal for 40 million unique visitors each day. This large volume of traffic is not a highly selected internet population because the Yahoo! Front Page provides up-to-date content intended to appeal to most demographics. Further, many new computers come pre-installed with www.yahoo.com as the default homepage. The ads on the Yahoo! Front Page provide a perfect opportunity to study frequency due to their single display location, their stable viewing population, their lack of competing ad units, and, most importantly, their unique sales strategy which creates a natural experiment.

Yahoo! Front Page advertisements are sold as "roadblocks," where all visitors to the page on a specific date are shown ads from one exclusive advertiser, or as "splits," where an advertiser purchases all display ad impressions delivered to visitors that arrive on an even second or an odd second. In this manner, each advertiser pays for only half of that day's display ad inventory but reaches 77% of the visitors with at least one impression, thus balancing reach and frequency. Importantly, the Yahoo! Front Page ad server ignores the identity of the user when deciding which ad to serve. Provided that whether a visit occurs in an even or an odd second is random, ad delivery is essentially a coin toss on "split" days and, hence, varies exogenously. This randomness of individuals' arrival time allows me to measure the effects of frequency on days where two advertisers each purchase a "split." On these days, for instance, individuals who visit the Front Page ten times see anywhere between zero and ten impressions from the "even-second" advertiser and the complement of their ten impressions from the "odd-second" advertiser. Consequently, the number of "even-second" ads seen is distributed binomially with the probability of seeing each advertiser's ad being one-half and the number of Bernoulli trials being the total impressions seen.

Figure 1-2 illustrates the empirical distribution of the number of one advertiser's impressions seen by users who visited the Yahoo! Front Page and saw ten ad impressions total between the two "split" advertisers during that day. As expected, the distribution is binomial with the probability of seeing one advertiser's ads or the other's being one-half. By comparing the clicker rates across groups of users who *randomly* saw different numbers of each advertiser's ads, I can measure the causal effects of frequency. Focusing on this random variation is important because unobserved heterogeneity across users of differing browsing intensities is correlated with the number of ads seen, confounding estimates of the impact of the ads. Simply stated, people who use the internet more see more ads and may differ in how they interact with and respond to the ads, relative to those who use the internet less. Specifically, individuals who frequent the Yahoo! Front Page may be more or less prone to click on an ad or sign up for an advertised online service than those who only visit occasionally. Using this natural experiment, I can hold the number of visits to the page fixed and observe responsiveness to the ads under different numbers of exposures.

In addition to the clean variation in numbers of exposures for users of a given type, another nice feature of the online environment is the easy access to responses to the ads. The two forms of response to the ads that I analyze are clicks and conversions such as new account sign-ups. Both ad clicks for 30 campaigns and new account sign-ups for one advertiser are available at the individual level and can be readily attached to the exposure data. These outcomes provide signals of the immediate impact of the display ads in the form of both interaction with the ad on the page and, subsequently, with the advertiser's online storefront.

III. Model

I present a simple semiparametric model for the effects of frequency on behavior which allows heterogeneity. I first consider model restrictions best suited to clicking behavior and then adapt those restrictions to accommodate additional outcomes, such as new account sign-ups.

A. Click Model

Let θ denote an individual's browsing type which is equal to the total number of ad impressions seen from the two advertisers. The browsing type, θ, is assumed to vary exogenously in the population and classifies users within the range from infrequent to frequent visitors, commonly referred to as light and heavy users, respectively. I would like to estimate the probability that an individual i of type θ clicks on the ads at least once when shown f impressions in a given advertising campaign c. In the econometric analysis, I assume that the clicking processed can be modeled as

$$\text{Clicked}_{ic} = a_c \cdot b_c(\theta_{ic}) \cdot h_c(f_{ic}) + \varepsilon_{ic}.$$

Clicked_{ic} is an indicator for whether the i^{th} person clicked at least once during the c^{th} campaign, a_c is the c^{th} campaign's scaling constant, θ_{ic} is the number of impressions eligible to the individual, and f_{ic} is the number of impressions that were delivered. The function $b_c(\cdot)$ defines the average heterogeneity in clicking propensity across individuals of different browsing types for the c^{th} campaign. The function $h_c(\cdot)$ defines how the probability of clicking varies with the number of ads delivered to the individual during that day. Finally, ε_{ic} is the error term which is independent of θ_{ic} and f_{ic}. In words, this independence assumes that the probability of clicking is only correlated with θ_{ic} and f_{ic} through the functions $b_c(\cdot)$ and $h_c(\cdot)$. This assumption is at the core of the exogenous variation induced by the coin-flip ad delivery that, for a given θ_{ic}, I observe a variety of values for f_{ic} which allows me to identify the causal relative effects of frequency function, $h_c(\cdot)$, separately from $b_c(\cdot)$.

This model defines a frequency response function for each campaign, $h_c(f)$, which explains how changing the number of ads influences the relative willingness of an individual to click.[3] Individual heterogeneity across browsing types scales this function up or down, depending on $b_c(\theta)$. Because the model is multiplicatively separable, both $b_c(\cdot)$ and $h_c(\cdot)$ are only separately identified up to scale. This means that $h_c(\cdot)$ is only a relative comparison of frequency given the scaling function $b_c(\cdot)$ and vice versa. Because both functions are only identified up to scale, I normalize $b_c(\cdot)$ with respect to $\theta=1$—the most precisely estimated parameter due to the low average frequency of the Yahoo! Front Page split campaigns—and $h_c(\cdot)$ with respect to f $=5$—an intermediate number of ads which accentuates the concavity or linearity from showing fewer or more than five impressions. Explicitly, these normalizations are $b_c(1)=1$ and $h_c(5)=5$. Further, since users cannot click without seeing an ad, I impose a location restriction of $h_c(0)=0$.

[3] The multiplicative separability assumption does not allow for different browsing types to have different relative frequency effect functions. This simplifying assumption allows us to define the response function for all frequencies observed and for all observed types.

These restrictions imply that the scaling factor a_c is equal to one-fifth of the clicker rate of users of type θ=1 who have seen f=5 ads for the c^{th} campaign.[4]

In order to estimate the model, I use nonlinear least squares, allowing for fully nonparametric specifications of $b_c(\cdot)$ and $h_c(\cdot)$. I use a combination of dummy variables and basis splines to permit adequate flexibility of the estimates for small values of the functions' arguments and to increase precision for larger values. Specifically, I restrict θ to natural numbers from 1 to 120. For $b_c(\cdot)$, dummy variables are included for θ=1 to 20, and b-splines span values of θ greater than 20 and up to 120. Regarding $h_c(\cdot)$, I include dummies ranging from f=0 to 15 and b-splines beyond 15 up to 50. Knots were generally equally separated by intervals of five, although some were omitted for larger values due to the increasing sparseness of the data.[5] The first-order splines I use allow $b_c(\cdot)$ and $h_c(\cdot)$ to be reasonably smooth functions and simplify nonlinear estimation relative to including additional weakly identified parameters for higher-order spline terms which are strongly collinear with the lower-order terms.

B. Conversion Model Restrictions

In the click model, I restricted $h_c(f)$ to go through the origin to impose the assumption that the click outcome cannot occur without the user viewing an ad. I now relax this assumption to model the baseline of other outcomes when no ads are seen. By allowing h(0) to differ from zero, the model can measure the effects of advertising frequency on other outcomes of interest to advertisers such as brand queries on Yahoo! Search, page views or new accounts on the advertiser's website, or online and offline sales. I focus on new account sign-ups:

$$(\text{New Account Sign-ups})_{ic} = a_c \cdot b_c(\theta_{ic}) \cdot h_c(f_{ic}) + \varepsilon_{ic}.$$

Note that this is equivalent to the click model, except now I impose $h_c(0)=1$ as the location and scale restriction. In this case all relative frequency effects are no longer in terms of the impact of five impressions (recall that $h_c(5)=5$ and $h_c(0)=0$ were the original restrictions), but rather in terms of the baseline sign-up rate.

[4] Clearly, a_c is identified through the model, because users of type θ=1 have only seen at most f=1 ads. However, the assumption caters to visual representations of $b_c(\cdot)$ and $h_c(\cdot)$ as functions and not to the explicit interpretation of a_c. However, a_c gives a sense of the scale of the clicker rate on the campaign's ads.

[5] For $b(\cdot)$, the b-splines include a constant for θ>20 and then knots at 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, and 100. For $h(\cdot)$, the b-splines include a constant for f>15 and then knots at 15, 20, 25, 30, 35, 40, and 45.

Additionally, to more precisely identify the relative frequency effects for a single advertiser, I impose the restriction that both $b_c(\cdot)$ and $h_c(\cdot)$ be equal across that advertiser's similar campaigns. To avoid any biases that cross-campaign variation in outcome levels, I include a multiplicative scaling factor, a_c, for each campaign:

$$(\text{New Account Sign-ups})_{ic} = a_c \cdot b(\theta_{ic}) \cdot h(f_{ic}) + \varepsilon_{ic}.$$

I use this model to examine the effects of frequency on new account sign-ups in section VI.

As with the click model, I combine dummy variables and b-splines to estimate the model using nonlinear least squares using visitors whose ad type θ ranged from 1 to 100 and whose ad exposure varied between 0 and 40 impressions. For $b(\cdot)$, dummy variables are included for $\theta=1$ to 4, and b-splines span values of θ greater than 4 and up to 100. Regarding $h(\cdot)$, I include dummies ranging from $f=0$ to 2 and b-splines beyond 2 up to 40.[6] Again, the first-order splines simplify estimation while providing sufficient flexibility to measure the functions.

C. Model Discussion

A naïve model might compare users who see different numbers of ads without accounting for the intrinsic differences across those users, potentially introducing endogeneity. The purpose of this model is to accommodate the exogenous variation identified in the natural experiment. The two overarching components of the model are the introduction of $b(\theta)$ and the conditional independence of ε, given θ and f.

The $b(\cdot)$ function allows users of different browsing intensities to have different absolute levels of the outcomes, thus respecting their heterogeneity. However, a naïve model that sets $b(\theta)=1$ for all browsing types, θ, would reduce the model to the implicit assumption that the number of ads that each visitor sees is effectively random. However, if visitors of various types differ in other ways that affect their willingness to click, regression estimates of $h(\cdot)$ will be biased due to a violation of the conditional independence assumption.

Thus, the model hinges on ε being independent of the model. This rules out scenarios such as the following: visitor *i* shows up and clicks the first ad that they see for the advertiser and

[6] For $b(\cdot)$, the b-splines include a constant for θ>4 and knots at 4, 7, 14, 19, 25, 30, 35, 40, 45, 50, and 60. For $h(\cdot)$, the b-splines include a constant for f>2 and knots at 2, 4, 7, 10, 14, 19, and 25.

go to that website, never to return to the Yahoo! Front Page; however, they would have seen ten ads had we not shown that advertiser's ad. Hence, the model may be misspecified because it ignores that an individual's click may distract them from returning, so θ_{ic} may be mismeasured and f_{ic} may not be as exogenously controlled as I assume.[7] However, I expect this to matter more for lower numbers of impressions than for higher numbers of impressions. In short, θ_{ic} and f_{ic} could be endogenous to whether a user clicks.[8]

One interpretation consistent with the independence assumption is that user *i* decides to visit the Yahoo! Front Page θ_{ic} times in a given day at exogenously set times and squeezes in around that preset schedule any additional browsing that results from clicking on the ad. While this likely is not a perfect description of true behavior, it probably is a good approximation.[9] However, any frequency analysis which depends on user behavior for the delivery will be subject to this criticism as the response to the most recently delivered ad could influence whether the next ad is delivered at all. In spite of these challenges, the combination of the natural experiment and modeling assumptions to account for heterogeneity across visitors is a significant step toward accurately measuring the effects of advertising frequency on behaviors.

IV. Data

The results are computed using three primary datasets for a number of Yahoo! Front Page "split" advertisers: individual-level impression and click data, raw event-level data for impressions and clicks for a single day (two campaigns), and online account sign-up data during four campaigns.

The individual-level impression and click data was obtained for fifteen days during January, February, and March 2010 when two advertisers shared all traffic for the day, providing a total of 30 campaigns to study. For each of these 15 days, I observe each user's anonymous unique identifier along with the number of ad views, *f*, and ad clicks for both advertisers. For each individual on each day, this allows me to construct the browsing type variable, θ, which

[7] Future research will investigate the severity of these effects and consider approaches to address this potential flaw. One approach would be to examine the "click-stream" data. If the likelihood of returning to the Yahoo! Front Page depends upon which ad you were shown, then this would suggest that the browsing type, θ, may be mismeasured.
[8] It is not clear that an instrumental variables approach where varying an advertiser's share of voice across randomized treatment groups would necessarily identify function $h(\cdot)$ if individuals do not return after clicking. More research should investigate the empirical relevance of these endogeneity concerns.
[9] Future work will attempt to endogenize θ while accounting for time-of-day effects and other concerns which could impact the econometric model.

tells me, "How many ads could this user have seen today for each advertiser?" With this information, I can compare outcomes such as clicks and views for users of the same type, θ, who were randomly shown differing numbers of impressions, *f*. Further, at the individual level, I am able to determine how many of the 15 days analyzed the same user visits the Yahoo! Front Page.

The raw event-level data for ad impressions and clicks was taken from a single day in March for campaigns 27 and 28. For each campaign, this event-level data recorded whether the event was an ad impression or click, the timestamp of the view or click, and the unique identifier for the visitor that viewed or clicked. This data allows me to derive the distribution of time between impressions and compute the length of time clicks typically follow views.

Finally, the online account sign-up data comes from a conversion beacon service which Yahoo! provides to advertisers. The relevant data corresponds to four campaigns, one of which is part of the 30 campaigns in the individual-level impression and click data. The data indicates which unique user triggered the beacon and when—either the day of the campaign or the day following. This data allows us to learn how the effects of frequency extend beyond just clicks to how they influence the likelihood of conversion.

Summary statistics for the 30 campaigns in the primary analysis are presented in Table 1-1. Each of the 15 days, roughly 40 million unique visitors trafficked the Yahoo! Front Page, viewing an average of 4.6 ads per day, split between the two advertisers. Each advertiser delivered between 70 and 98 million impressions and attracted approximately 100,000 clickers per day, on average.[10] There is substantial heterogeneity in the response rates to different advertisers: some advertisers were able to attract clicks from as many as 1.36% of all visitors to the Yahoo! Front Page, while most others attracted between 0.10% and 0.30%. In total, I observe 1.2 billion visitor×campaigns which equate to 609 million visitor×days—two campaigns per day. These visitor×days included only 237 million unique visitors, or roughly 39% of 609 million. Of the unique visitors, 146 million visited only one of the 15 days, 7 million visited five days, 2 million visited ten days, and 2 million showed up all 15.

[10] I focus on clickers in order to be conservative—the same user clicking twice in one day is still just a single visitor that day. By eliminating duplicate clicks, any frequency estimates will necessarily be slightly downward biased because users that browse longer will only have their earliest clicks counted. While multiple-clickers might be more valuable than single-clickers, I eliminate all duplicate clicks and leave that question to future research. As can be seen by the small differences (~10%) between total number of clicks and clickers for each campaign in Table 1-1, multiple-clickers form a small share of the clickers, so this assumption should not greatly influence my results.

Figure 1-3 shows the highly-skewed distribution of browsing types, θ, on a typical day, day 5. That day, 90% of visitors saw fewer than 10 impressions for both split advertisers. While visitors saw an average of 4.7 ads that day, the skewed distribution places the median on the margin of 2 and 3 impressions. Under usual circumstances such skewed data might leave little hope of identifying any effects out as far as 20 or 30 impressions, however, the scale of the data produces a remarkable 5,780 unique users who were exposed to exactly 50 impressions (θ=50) on the Yahoo! Front Page that day. As such, the scale of the data allows me to obtain reasonably tight confidence intervals out as far as f=40 when estimating the models.[11]

The even-odd second alternation of advertisers induces a binomial distribution in the number of each advertiser's impressions delivered, f, for a given browsing type. For the 2.3 million users with browsing type of θ=5 on day 14, a comparison of the empirical distribution of delivered impressions and binomial(θ=5,p=0.5) yields a Pearson's χ^2 statistic of 4.9, which has a critical value of 11.1. This confirms our hypothesis of the delivery method. However, 3% of impressions that were registered with the same timestamp or delivered within one second of the previous ad were eliminated for the sake of the test.[12] As long as the duplicate impressions are essentially random or account for a negligible share of the total impressions, they should have little influence on the qualitative results the frequency analyses.

V. Analysis of Clicks

I estimate the marginal effects of display ad exposure on ad clicks. In particular, I assess the degree to which there are increasing, constant, or decreasing returns to scale for the ads shown to users for single-day campaigns shown on the Yahoo! Front Page. For each of the 30 campaigns, I estimate the click model described in section III.A, repeated here for convenience:

[11] Stronger smoothness assumptions could easily extend the estimates beyond f=50.

[12] An interesting note is that there is a difference between the ad serving decision and the logging of the ad-serve event. It turns out that ~20% of the ads that were supposedly served on an even second were actually logged on an odd second. This suggests that there may be an average latency of 200 milliseconds between the ad serving decision and the logging of the event. This means that even and odd seconds in the server logs did not necessarily indicate the same or different ads. Specifically, if two impressions for the same user had the same timestamp recorded between the two, it is impossible to determine from the data which impression came first! So, in order to remedy conformance to the binomial, I selected a random number to decide which of the two impressions should be removed and which should be kept. This trick, which might have introduced additional misspecification bias if I had been mistaken about the cause of the deviation from the binomial distribution, caused the empirical distribution to conform nicely to the predicted binomial theory as shown by the Pearson's χ^2 goodness-of-fit statistics presented.

$$\text{Clicked}_{ic} = a_c \cdot b_c(\theta_{ic}) \cdot h_c(f_{ic}) + \varepsilon_{ic}.$$

Most of the campaigns represent distinct advertisers, and several days separated campaigns from the same advertiser.[13] However, each Yahoo! Front Page campaign does not necessarily account for all ads being shown by the advertiser—I estimate the frequency effects of showing these ads at this online location, taking all other contemporaneous advertising for this advertiser as given.[14]

A. Effects of Frequency on Clicks

Figure 1-4 shows the estimate of the absolute frequency effect, $a_c \cdot h_c(f_{ic})$, for all 30 campaigns. Each campaign exhibits its own unique behavior with respect to the absolute click performance of the campaigns as well as to the relative shapes of the curves according to the number of ads seen. Some campaigns have very high clicker rates while others have relatively lower click-response to the ads. More interestingly, several campaigns appear to exhibit clicker rates which taper off with additional impressions while others appear to be virtually immune to a clicker-rate upper bound with each additional impression yielding nearly the same number of incremental clicks.

Figure 1-5 focuses only on the relative frequency effect, $h(f)$, and shows the relative performance from showing additional ads for all 30 campaigns. Note that the h(5)=5 scale normalization has been applied to each frequency estimate to allow for scale-free comparisons across campaigns. This normalization implies that the relative frequency estimates are in units of 20% of the impact of showing five impressions. Two aspects of the plot immediately stand out. First, several campaigns experience rapid wear-out, with one campaign's estimate showing h(1)=3, or roughly 60% of the impact of showing five ads occurs from the first display ad shown. These ads are seen and clicked within a small number of impressions, experiencing substantial decreasing returns to scale. Second, many campaigns experience relatively little wear-out, with h(20)=20, implying that four times the impact of showing five ads is achieved by showing

[13] Future research will investigate the impact of adjacent days: to what extent does the frequency consumed yesterday influence behaviors today? Due to nonlinear pricing of the "split" versus "roadblock" ads, the same advertiser showing ads on adjacent days occurs very infrequently, but a few examples do occur. Such research also could be extended to other repeat advertisers for non-adjacent days.

[14] For large media campaigns where the Yahoo! Front Page campaign is concurrent with a large amount of other online and offline advertising, rapid wear-out could be partially explained by exposure to the advertiser's message through different media on the same day.

twenty. For these advertisers, each incremental impression is just as potent as the last; the ads experience constant returns to scale.

Among the 30 campaigns, there are ten campaigns with *constant returns to scale* (h(20) in [18,21]), six with *mild wear-out* (h(20) in [15,18]), ten with *moderate wear-out* (h(20) in [12,15]), and four with *extreme wear-out* (h(20) in [8,12]). In words, the benefit of showing twenty impressions ranges from 1.7 to 4.1 times the impact of showing five impressions—from strongly decreasing returns to scale to constant returns to scale. The four campaigns in Figure 1-6 illustrate the varying degrees of relative frequency performance for the 30 campaigns. The top left panel provides an example of a campaign where the estimate of the relative frequency curve follows the Y=X line in the plot which coincides with constant returns to scale. The top right panel illustrates the *mild wear-out* as evidenced by the increasingly large deviation from the Y=X line for higher frequencies. The bottom left panel shows *moderate wear-out*, with an even larger deviation from constant returns. However, the bottom right panel shows a campaign where the wear-out is so extreme that an additional 35 ads are necessary to equal the impact of the first five.

None of the curves suggest negative impacts from showing too many impressions to users within a single day (commonly known as the "inverted U") because all marginal effects of ads on clicking are positive. However, if marginal willingness to click did turn negative, in a "hold-up" fashion, the advertiser may have already received the clicks that were expected from showing a smaller number of impressions earlier. Thus, heterogeneity in attitudinal effects could be both positive and negative as interest in clicking dwindles for some viewers, but the frequency estimates for clicks would be unable to illustrate the negative impact of overexposure to the ad. While the level of statistical precision does not rule out constant effects for relevant frequencies of several of the campaigns, additional attitudinal or direct-response (e.g., sales and sign-ups) outcomes should be used to investigate the risks of overexposure to the ads in future work.

B. Browsing Type Heterogeneity

The reasoning behind my use of this natural experiment to identify the effects of frequency is because I expect browsing heterogeneity to impact the results. I find that the effects of browsing heterogeneity are important because the relationship between browsing types and users' propensity to click varies from campaign to campaign.

The browsing type heterogeneity function, b(θ), lets me identify differences in light and heavy users' willingness to click on ads for each campaign. Figure 1-7 shows the estimated functions for each of the 30 campaigns analyzed, where each has been normalized such that b(1)=1 (i.e., relative to users who only visit the Yahoo! Front Page once during the day). In the absence of heterogeneity, meaning, if the number of visits to the Yahoo! Front Page was effectively randomly determined, the entire function should be constant and equal to one. However, most of the campaigns exhibit substantial variation in the relative propensities to click with some estimates showing decreasing propensities to click in browsing intensity and others showing increasing click propensities.

Again, I examine four campaigns whose curves are representative of the range of the browsing type heterogeneity estimates which I have divided into four[15] categories presented in Figure 1-8. The 30 campaigns are divided into fifteen for which light users respond more (top left panel; b(50) in [0.45,0.95]), eight for which responsiveness is roughly constant across browsing types (top right panel; b(50) in [0.95,1.25]), five for which heavy users respond more (bottom left panel; h(50) in [1.25,1.75]), and two for which heavy users respond much more than light users (bottom right panel; h(50) in [1.75,2.25]). Several campaigns experience one-half the responsiveness between light and heavy users as illustrated in the top left panel where users with browsing types of 100 are half as likely to click on the ad, holding the relative frequency effects and CTR constant. Others experience near-constant responsiveness for all users ranging from light to heavy users as shown in the top right panel of Figure 1-8—for these campaigns, browsing-type heterogeneity appears to not matter much. Finally, as illustrated in the bottom two panels, some campaigns experience greater response from heavier users, with some campaigns experiencing as much as 2.5 times the clicking responses from heavy users relative to the light users (θ=100 versus θ=1). Nevertheless, 90% of visitors to the Yahoo! Front Page are of browsing type 10 or less, and a closer look at this important portion of the browsing type curves in Figure 1-9 reveals that the curve is greater than one for virtually all campaigns. Therefore, users who visit between two and ten times are, on average, more willing to click than users who visit only once, controlling for frequency effects.

[15] The definitions of these four categories are restrictive and one-dimensional, only allowing for a relative comparison between users of browsing type 1 and 50. A few of the campaigns defy meaningful classification in this way because their curves rise sharply for the users of types one, two, or three but then slowly decline in users' browsing type.

Ignoring this widely-varying, cross-campaign heterogeneity in clicking propensities across users of different browsing intensities induces bias into frequency estimates. Further, no simple correction, such as multiplying by a constant, can account for user heterogeneity when estimating the impact of frequency on clicking. The browsing type heterogeneity function is, perhaps, the simplest way to reliably account for user heterogeneity in this application.

The browsing type heterogeneity function is a composition of two factors: differences in the level of intrinsic interest in the ad's offering and in the advertiser's stock (Palda 1964, Mann 1975). Interest in the advertiser's offering may be correlated with browsing intensity. For example, advertisements for online games might appeal to heavier internet users while ads for consumer packaged goods might appeal to lighter internet users who also frequency shop offline. That said, if an advertiser frequently advertises online, heavier users will experience proportionally more exposure to the ads and, hence, a larger stock of the ads. This could cause the marginal responsiveness of heavier users to be much lower, especially if they have already responded to the company's offering. Consequently, their willingness to click may be much lower—not because they were not good candidates for the ads, but because they had already responded. Regardless, the shape of b(θ) can signal whether an advertiser's budget is better spent on light or heavy users of the Yahoo! Front Page because the relative differences in responsiveness translate into absolute differences, after controlling for differences in expected wear-out.

C. Estimation Ignoring Heterogeneity

After accounting for browsing heterogeneity when estimating the impact of frequency, it is possible to impose the constraint of b(θ)=1 for all θ and obtain the naïve estimates. Imposing this constraint is equivalent to asking the observational question: do visitors who see different numbers of ads click differently? This question directly contrasts with the causal frequency question of interest: how does the same visitor's behavior vary when shown different numbers of ads? I compare the estimates obtained accounting for heterogeneity with naïve estimates which assume that the number of ads seen is exogenous and, hence, that b(θ)=1 for all θ.

Four of the campaigns, shown in Figure 1-10, illustrate the range of conclusions resulting from naively assuming that all internet users are the same. In the top left panel, one campaign observed to be experiencing constant returns to scale when correctly accounting for the heterogeneity shows significantly decreasing returns to scale when ignoring the differences in browsing types. Many other campaigns show this effect, but to a lesser extent as in the top right panel. Still, several campaigns would have realized nearly the same results using either method—this is the case for the bottom left panel—while other campaigns actually wore-out faster than the naïve estimates would suggest. This means that, by ignoring browsing type differences, campaigns that are experiencing constant returns to scale could be incorrectly designated as experiencing decreasing return to scale while other campaigns that are wearing out fast could be miscategorized as experiencing wear-out more slowly. Both of these mistakes are potentially costly.

For the 30 campaigns analyzed, biases resulting in the overstatement of wear-out are severe. Figure 1-11 shows the distribution of bias ratios of h(40) for the 30 campaigns. Only four campaigns experienced understated estimates of wear-out. The remaining 26 campaigns' wear-out was overstated by the naïve estimator with 16 campaigns experiencing biases ranging between -32% and -62%. This significant overstatement of wear-out made by the naïve model highlights the importance of accounting for user heterogeneity when estimating the effects of online advertising.

Finally, I examine an assertion frequently heard in industry (Chang and Thorson 2004): there is a convex portion of the frequency response function, h(*f*), indicating that there are increasing returns to scale, or ad "synergy" to showing multiple ads because exposure to a few impressions is necessary for visitors to notice the ad's message and overcome their reluctance to click. I examine the frequency response functions more closely to assess this claim by comparing the relative frequency effects of showing one ad versus two for the 30 campaigns. Figure 1-12 shows a comparison of h(2) – 2*h(1) for the unconstrained and constrained (naïve) estimators. This is a simple comparison asking, "Is the likelihood of clicking on the second ad more than twice the likelihood of clicking on the first?" The unconstrained estimator only finds two campaigns (in the first quadrant) that might[16] exhibit weak synergy from showing the second ad, with the largest of the two representing a 1.0% increase in h(2) from synergy. However, the naïve estimator designates half of the 30 campaigns (first and second quadrants) as exhibiting synergy with 3-6% boosts in responsiveness to the second ad, h(2), for 13 campaigns.

[16] Note that confidence intervals for the unconstrained frequency estimates are much wider than for the naïve estimator which makes use of the endogenous variation as well. The confidence intervals do not rule out the possibility of no synergy.

The naïve estimates show synergy that the estimates which account for differences in browsing types do not. The severity of the false synergy estimated here is not great, mostly due to the untargeted nature of these ads. However, I would expect to find a much larger spurious estimate for highly targeted ads where the synergy actually derives mechanically from the ad-targeting algorithms selecting users. That said, an absence of synergy is not necessarily bad—economically, the benefits of higher frequency are achieved as long as the marginal benefit of the additional clicks exceeds the marginal cost of delivering additional impressions.

D. Heterogeneity and Wear-Out

The wide range of varying ad performance begs an explanation. Perhaps some ads are better at being noticed. Alternatively, some ads may be more prone to wear out simply because they fail to appeal to a general audience. There are tradeoffs between targeted messages that resonate with a particular segment of the population and fail to connect with the remaining viewers and a broader message that will not expire. It could be that the information conveyed by the ad is dated or time-sensitive such that visitors see the ads, find out the information, and move on after one impression. What creates such differences in frequency behavior across ads, and should advertisers be more concerned about the ads that wear out fast or those that fail to wear out?

Does the freshness of the creative matter? A number of the advertisers who experience decreasing returns to scale also experience some of the highest CTRs. From a basic review of the ad creatives, the content of these ads tended to be newer or more time-specific. In particular, a number of the ads experiencing the greatest wear-out were associated with new product releases (3 campaigns), new television episode (3 campaigns) or video releases (4 campaigns), or other time-sensitive or novel content (9 campaigns). The advertisers who experience constant returns to scale for their advertising exposure tended to display similar creatives (6 campaigns) or represent well-known brands for which there would arguably be relatively little new information coming from the ad (5 campaigns), other than a new opportunity to be reminded about the brand or product.

The wide variation in wear-out across advertisers may raise concerns about advertising effectiveness. However, advertisers should only worry about wear-out if the marginal effectiveness of the ad becomes sufficiently low for it to be unprofitable to advertise—not merely if an ad becomes less effective beyond the first exposure.[17] There is a trade-off between showing more ads to the best candidates and reaching marginal consumers who are less likely to respond. If the ads wear out slowly, then advertisers should not worry about reach, but should only target the most responsive subpopulation. If the ads wear out quickly, then advertisers should trade off the cost-ineffective frequency to reach more people, up to where the benefits to reach equal the benefits to frequency for the inframarginal reach and frequency.

VI. Analysis of Conversions

Campaign 10 was run by an advertiser with whom Yahoo! has a data-sharing partnership for anonymized conversion tracking. As a result, I can match users that were shown the advertiser's ads on the Yahoo! Front Page and were shown a confirmation page upon completing the sign-up process for a new account on the advertiser's website. This conversion data lets me measure the impact of frequency on this important outcome.[18] I use data from four campaigns run on four days.[19] Roughly 700 new account sign-ups occurred among the 40 million ad viewers each of the four days.

I estimate the conversion model from section III.B,

$$(\text{New Account Sign-ups})_{ic} = a_c \cdot b(\theta_{ic}) \cdot h(f_{ic}) + \varepsilon_{ic},$$

where h(·) is no longer required to go through the origin, but through h(0)=1. The relative frequency effects are displayed in Figure 1-13. In this figure, the relative frequency effects for sign-ups are in terms of the baseline sign-up rate when no ads are shown. The plot shows a large and positive impact of frequency from showing even as many as 20 impressions. The frequency impact on clicks for campaign 10 in Figure 1-15 are similar to those for sign-ups—both curves show relatively little wear-out after users view a large number of impressions in a single day. I

[17] I forgo discussing investment in new creatives for the sake of brevity. While ongoing investment in fresh new messages is encouraged by researchers such as Weilbacher (1970), I limit the advertiser's choice set "to whom should I target the ads" and "how many should be shown."

[18] It should be noted that the advertiser does not claim this to be the only important outcome that they are interested in for the campaign. Further, the data from the conversion beacon being analyzed here only accounts for a fraction of the total number of sign-ups registered by users due to browser cookie deletion. As such, the purpose of this analysis is primarily qualitative: does advertising frequency influence other important conversion outcomes such as account sign-ups?

[19] Each of the four campaigns was separated by at least a week from the other three campaigns.

compare the average effect of viewing twenty impressions on conversions with the effect of viewing only one. The estimated effect of one impression is 31% of the baseline and of twenty impressions is 627%, a ratio of 19.95.[20] Using the estimates of the frequency effects for clicks, the average effect of viewing the ads was 1.02 and 15.05 for one and twenty impressions, respectively, for a ratio of 14.79. These two ratios are statistically indistinguishable, indicating similar frequency effects for sign-ups as for clicks.

For the day following each campaign, the frequency impact on sign-ups was positive for small numbers of impressions, although the effects were statistically insignificant. However, the long-run impact of advertising on sign-ups is difficult to identify because the effect likely decays over time, reducing the signal-to-noise in estimation which is barely surmountable for the same day the ads are shown. Future research will pursue changes in both the estimation strategy and ad delivery that will improve the precision of these estimates.

The browsing heterogeneity curve in Figure 1-14 clearly deviates from the constant line of $b(\theta)=1$, with heavy users being between two and three times more responsive to the advertiser's offering than users who only visit the Yahoo! Front Page once. Ignoring this browsing type heterogeneity would produce naïve frequency estimates that significantly overstate the impact of frequency on sign-ups for the four campaigns. The estimate of the causal effects of frequency on sign-up rates for the advertiser's four campaigns provides a much more reliable estimate of the impact of showing more ads to the same user.

VII. Conclusion

In this paper, I measured the impact of display advertising frequency for 30 campaigns shown on the Yahoo! Front Page between January and March 2010. The main finding for these campaigns is that the effect of display ad frequency on clicking behavior exhibits large heterogeneity. Some advertisers find little decline in clicks to the marginal ad, even for as many as 30 to 50 impressions, while other advertisers experience rapidly decreasing clicks for showing even as few as two or three impressions. For users' click responses to the 30 campaigns, ten campaigns experience constant returns to scale, six experience mild wear-out, ten experience moderate wear-out, and four experience extreme wear-out. For one advertiser, I can estimate the

[20] Note that this ratio is very imprecisely measured due to the volatility of the estimate of $h(\cdot)$ for f=20 as seen in Figure 1-13.

impact of frequency on new account sign-ups on the advertiser's website and find similar mild wear-out frequency effects for both sign-ups and clicks for that advertiser.

Carefully examining the frequency estimates, I find no convincing evidence for increasing returns to scale, commonly known as "ad synergy." However, naïve estimates that ignore correlation of unobserved heterogeneity with the number of ads find false synergy for half of the 30 campaigns. Conversely, the naïve estimator significantly overstates the wear-out of most of the 30 campaigns and understates the wear-out for a few.

Whether or not an ad faces a declining marginal response rate, advertisers should show precisely the number of ads that equate the marginal benefit reaped from the last ad shown with its marginal cost. By better understanding the effects of frequency as well as the accessibility of impressions to the target audience, better decisions can be made to deliver the right number of ads to the most responsive audience. Rather than merely "increasing reach" as many practitioners seek, this knowledge can help advertisers "reach the right customers."

Future research will bifurcate: additional work should focus on examining the causes of variation in the frequency curves and on reliably measuring the effects of frequency in a more circumstances. Regarding the former, decomposing the impact of frequency by demographic as well as technographic measures such as age, gender, and browsing behaviors may lead to valuable insights as to what influences visitors' responsiveness to the advertising.

Regarding the latter, the model presented in this paper has been applied under unique circumstances with a special form of exogenous variation. Alternative models which make use of time-series variation should be evaluated to assess their reliability at measuring the effects of frequency. One such model which I leave to future research[21] asks, "How many clicks occur on the n^{th} ad that is delivered?" For example, for users that see three ads, what is the expected CTR on the first, second, and third impressions? If time-varying effects are negligible, this model could approximate the experimental results without the natural experiment. However, attributing non-click outcomes to specific views is notoriously challenging in the absence of an

[21] A preliminary analysis using the raw impression and click logs for campaigns 27 and 28 suggests that examining the cumulative click behaviors could be a reasonably good approximation to the results of the natural experiment—after accounting for heterogeneity of browsing type, of course. The primary shortcoming of the analysis, however, is a lack of robustness to a variety of time-dependent concerns such as time-variant click propensities. For example, the average time at which the second ad is delivered is always later in the day than the average time for the first ad, whereas a randomly-delivered single ad can have the average characteristics of the first and second ads for users who are delivered two ads total if users are randomly delivered one or two ads. Future research will evaluate the robustness of this model's performance with highly targeted ads.

experiment.[22] For these reasons, I have concentrated on a model which uses the exogenously-varying portion of the data in the natural experiment to obtain credible estimates of the effects of frequency.

VIII. References

Campbell, Margaret C. and Kevin Lane Keller. 2003. "Brand Familiarity and Advertising Repetition Effects." *Journal of Consumer Research*, 30(2): 292-304.

Carlon, M. and M. Hislop. 2001. "5 Exposure Plateau for Building Message Online." *Beyond The Click: Insights from Online Advertising Research*, 1(8). Retrieved October 28, 2001 from www.dynamiclogic.com/beyond_1_8.html

Chang, Yuhmiin, and Esther Thorson. 2004. "Television and Web Advertising Synergies," *Journal of Advertising*, 33(2): 75-84.

Chatterjee, Patrali, Donna L. Hoffman and Thomas P. Novak. 2003. "Modeling the Clickstream: Implications for Web-Based Advertising Efforts," *Marketing Science*, 22(4): 520-541.

Chiou, Lesley and Catherine Tucker. 2010. "How does Pharmaceutical Advertising affect Consumer Search?" Working paper. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=IIOC2010&amp;paper_id=514

Craig, C.S., B. Sternthal, and C. Leavitt. 1976. "Advertising wearout: An experimental analysis." *Journal of Marketing Research*, 13(4): 365-372.

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

Dreze, X and F. X. Hussherr. 2003. "Internet advertising: Is anybody watching?" *Journal of Interactive Marketing*, 17(4): 8-23.

Havlena, William, Robert Cardarelli, and Michelle de Montigny. 2007. "Quantifying the Isolated and Synergistic Effects of Exposure Frequency for TV, Print, and Internet Advertising." *Journal of Advertising Research*, 47(3): 215-221.

Kameya, Alison and Katherine Zmija. 2002. "What Makes Online Advertising Effective?" Michigan State University.

Lewis, Randall A. and Taylor A. Schreiner. 2010. "Can Online Display Advertising Attract New Customers? Measuring an Advertiser's New Accounts with a Large-Scale Experiment on Yahoo!" Working paper, Yahoo! Research.

Lewis, Randall A. and David H. Reiley. 2010a. "Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via a Controlled Experiment on Yahoo!" Working Paper, Yahoo! Research.

Lewis, Randall A. and David H. Reiley. 2010b. "When the Marginal Ad Isn't Marginal: Equating Marginal Benefit and Marginal Cost in a Large-Scale Advertising Experiment on Yahoo!" Working Paper, Yahoo! Research.

[22] See Lewis and Schreiner (2010) for more discussion about attribution of the effects of advertising to publishers.

Mann, Don H. 1975. "Optimal Theoretic Advertising Stock Models: A Generalization Incorporating the Effects of Delayed Response from Promotional Expenditure." *Management Science*, 21(7): 823-832.

Morgan Stanley Dean Witter. 2001. "Does Internet Advertising Really Work?" Retrieved October 20, 2001 from www.morganstanley.com/techresearch/

Palda, Kristian S. 1964. "The Measurement of Cumulative Advertising Effects." Prentice-Hall: Englewood Cliffs, N.J.

Pechmann, Cornelia and David W. Stewart, 1990. "Advertising Repetition: A Critical Review of Wear-In and Wear-Out," Marketing Science Institute. pp. 90-106.

PriceWaterhouseCoopers, L.L.C. "IAB Internet Advertising Revenue Report 2009." http://www.iab.net/insights_research/1357.

Robinson, Helen, Anna Wysocka, and Chris Hand. 2007. "Internet advertising effectiveness: the effect of design on click-through rates for banner ads." *International Journal of Advertising*, 27(4): 527-542.

Simon, Julian L. 1979. "What Do Zielske's Real Data Really Show About Pulsing," *Journal of Marketing Research*, 16(3): 415-42.

Tellis, Gerard J. 1988. "Advertising Exposure, Loyalty, and Brand Purchase: A Two-Stage Model of Choice." *Journal of Marketing Research*, 25(2): 134-144.

Tellis, Gerard J. 1997. "Effective Frequency: One Exposure or Three Factors?" *Journal of Advertising Research*, 37(4): 75-80.

Weilbacher, W. M. 1970. "What Happens to Advertisements When They Grow Up," *The Public Opinion Quarterly*, Vol. 34(2): 216-223.

Zielske, Hubert A. 1959. "The Remembering and Forgetting of Advertising," *Journal of Marketing*, 23(1): 239-43.

IX. Appendix

I include a few additional details confirming the quality of the identification strategy, the complete results for all 30 campaigns estimated, and additional figures assessing the goodness-of-fit of the model.

While I claim that the ads alternate on even and odd seconds for the Yahoo! Front Page split campaigns, seeing is believing: Figure 1-16 shows the nearly identical 5-minute ad delivery rates over the course of a day (US Eastern Time) for two advertisers splitting the Yahoo! Front Page on day 14. This confirms the identification strategy while providing insight into the scale of the engineering required—at its daily peak, nearly 500,000 impressions are delivered per five-minute interval, while in the nightly trough 50,000 impressions are delivered.

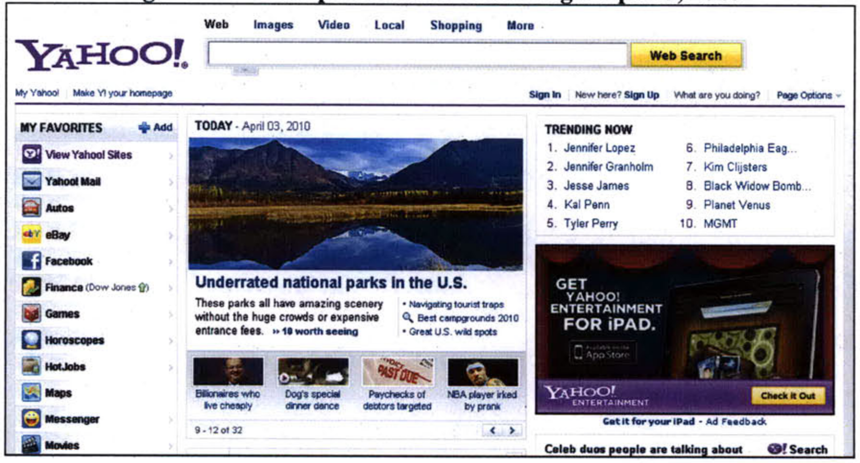
The model was estimated on all 30 campaigns. In order to examine each campaign individually and understand the statistical precision of the relative frequency estimates, I present each campaign's estimate of $h_c(f)$ with 95% confidence bounds in Figure 1-17 for the interval from f=1 through f=20. This is the most relevant interval for these campaigns. I also present estimates of $b_c(\theta)$ for each campaign separately in Figure 1-18 to clearly illustrate the level of confidence of the estimates and for easy comparison with the relative frequency effects shown in Figure 1-17. Additionally, I include comparisons between estimates of $h_c(f)$ for the naïve constrained model where $b_c(\theta)=1$ for all θ and for the unconstrained model in Figure 1-19 for f=1 through f=50 to highlight the differences between the two estimators for larger frequencies which are more telling about wear-out.

Finally, I perform a simple, visual specification test to validate the relative frequency estimates. In this test, I obtain the relative frequency slope of the binomially-varying frequency for each browsing type from 1 to 40. I do this by estimating an ordinary least squares (OLS) regression of whether the visitor clicked on the number of impressions seen for each browsing type.[23] I then plot the line perpendicular to the slope of the OLS estimate through the relative frequency estimate at the expected frequency seen by visitors of that particular browsing type. This visual test is presented in Figure 1-20 for campaign 1. Note that the arrows are roughly orthogonal to the relative frequency estimate and, as the confidence interval widens as the data becomes thinned out, the stability of each arrow's direction declines. The test illustrates that the model is a reasonably good fit by showing that the fully nonparametric slope estimates for each type are roughly parallel to the slope of the relative frequency estimates. If the multiplicative separability assumption was violated, this would not be the case. Instead, the arrows would vary systematically from being orthogonal to the relative frequency estimates. Figure 1-21 provides an example of how the multiplicatively separable model might systematically differ from a nonseparable model.

[23] Specifically, I estimate a local slope of the frequency curve about the expected frequency, $f=\frac{1}{2}\theta$. For each θ, the regression constrains observations to $\theta=\theta_0$ and $\frac{1}{2}\theta - 2 \leq f \leq \frac{1}{2}\theta + 2$. Thus, for each browsing type I can observe the average marginal effect of the ad about the binomial distribution's peak where the data is mostly concentrated and the average marginal effect is most precisely estimated.

X. Figures and Tables

Figure 1-1 - Example of Yahoo Front Page: April 3, 2010



TODAY - April 03, 2010

Underrated national parks in the U.S.

These parks all have amazing scenery without the huge crowds or expensive entrance fees. » 10 worth seeing

- Navigating tourist traps
- Best campgrounds 2010
- Great U.S. wild spots

Billionaires who live cheaply

Dog's special dinner dance

Paychecks of debtors targeted

NBA player irked by prank

9 - 12 of 32

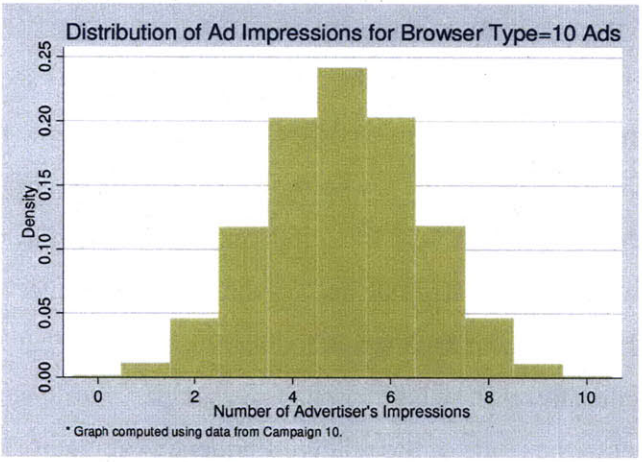
TRENDING NOW

1. Jennifer Lopez	6. Philadelphia Eag...
2. Jennifer Granholm	7. Kim Clijsters
3. Jesse James	8. Black Widow Bomb...
4. Kal Penn	9. Planet Venus
5. Tyler Perry	10. MGMT

Get it for your iPad - Ad Feedback

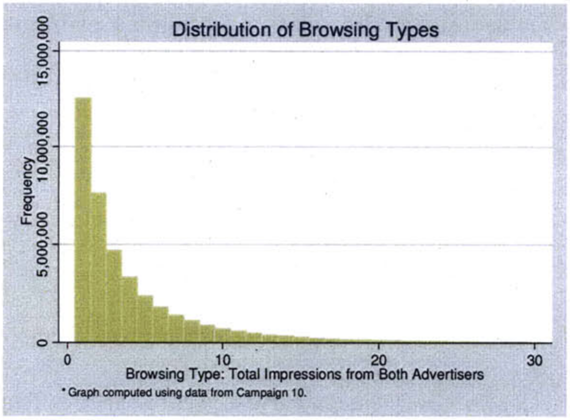
Celeb duos people are talking about

Figure 1-2 - Ad delivery produces a binomial distribution for an advertiser's impressions.



* Graph computed using data from Campaign 10.

Figure 1-3 - User browsing types are concentrated below ten Yahoo! Front Page ad impressions in a day.



* Graph computed using data from Campaign 10.

Table 1-1 - Summary statistics for 30 Yahoo! Front Page "split" campaigns over 15 days.

Campaign		Impressions	Visitors	Imp./Visitor		Clicks	Clickers	CTR	% Clickers
Day 1	1	98,177,379	41,221,308	4.76	2.38	51,653	46,010	0.05%	0.11%
	2	98,155,857			2.38	62,729	52,763	0.06%	0.13%
Day 2	3	98,359,566	41,467,202	4.75	2.37	52,667	47,316	0.05%	0.11%
	4	98,428,392			2.37	117,716	98,759	0.12%	0.24%
Day 3	5	98,718,046	41,993,478	4.70	2.35	50,004	44,633	0.05%	0.11%
	6	98,630,270			2.35	156,878	145,943	0.16%	0.35%
Day 4	7	96,484,810	41,770,305	4.62	2.31	130,056	108,510	0.13%	0.26%
	8	96,576,838			2.31	117,795	105,606	0.12%	0.25%
Day 5	9	94,700,877	40,611,100	4.67	2.33	54,142	48,665	0.06%	0.12%
	10	94,809,456			2.33	46,790	42,114	0.05%	0.10%
Day 6	11	78,106,177	32,700,667	4.78	2.39	61,260	53,084	0.08%	0.16%
	12	78,157,493			2.39	27,783	25,415	0.04%	0.08%
Day 7	13	91,620,276	41,811,276	4.38	2.19	284,565	253,503	0.31%	0.61%
	14	91,655,105			2.19	139,295	116,925	0.15%	0.28%
Day 8	15	94,915,890	42,503,753	4.47	2.23	47,617	42,937	0.05%	0.10%
	16	94,924,172			2.23	49,677	37,316	0.05%	0.09%
Day 9	17	70,325,235	31,925,737	4.41	2.20	104,522	87,509	0.15%	0.27%
	18	70,337,837			2.20	28,911	26,785	0.04%	0.08%
Day 10	19	97,637,885	43,040,340	4.54	2.27	50,334	44,405	0.05%	0.10%
	20	97,596,427			2.27	96,141	88,400	0.10%	0.21%
Day 11	21	95,612,568	42,513,766	4.49	2.25	179,173	158,226	0.19%	0.37%
	22	95,248,995			2.24	48,762	44,585	0.05%	0.10%
Day 12	23	98,032,549	43,175,570	4.54	2.27	486,067	451,975	0.50%	1.05%
	24	98,053,868			2.27	50,313	44,571	0.05%	0.10%
Day 13	25	93,770,718	42,221,664	4.44	2.22	659,665	575,327	0.70%	1.36%
	26	93,816,345			2.22	89,623	80,049	0.10%	0.19%
Day 14	27	88,367,477	40,227,548	4.39	2.20	69,041	62,514	0.08%	0.16%
	28	88,299,108			2.19	75,584	68,364	0.09%	0.17%
Day 15	29	94,742,664	42,092,579	4.50	2.25	46,686	41,691	0.05%	0.10%
	30	94,714,655			2.25	71,779	66,901	0.08%	0.16%
Totals & Avgs.:		2,778,976,935	609,276,293	4.56	2.28	3,507,228	3,110,801	0.13%	0.51%

Figure 1-4 - Absolute frequency effects for all 30 campaigns.

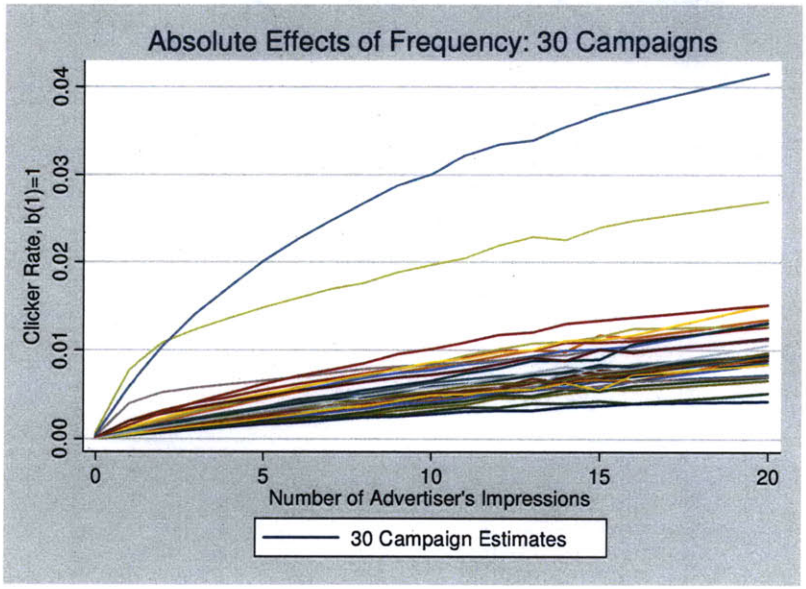


Figure 1-5 - Relative effects of frequency for all 30 campaigns.

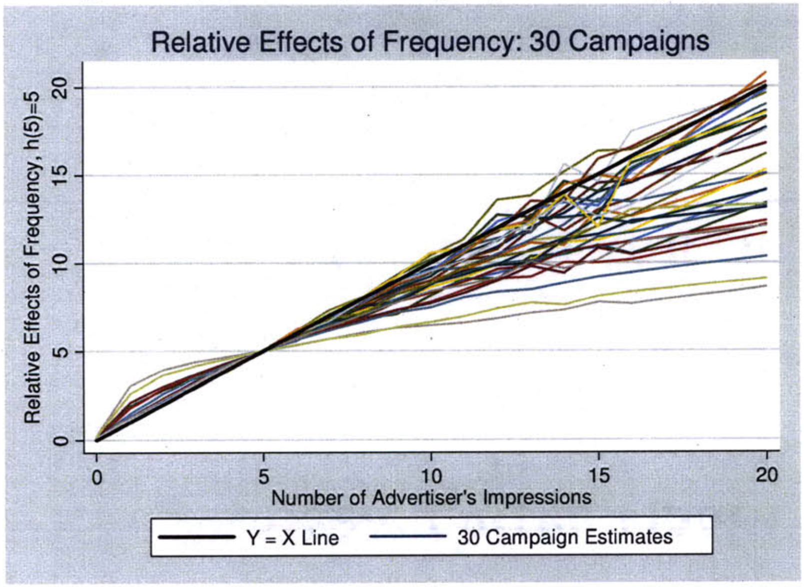


Figure 1-6 - Four examples highlight the heterogeneous relative frequency response across campaigns.

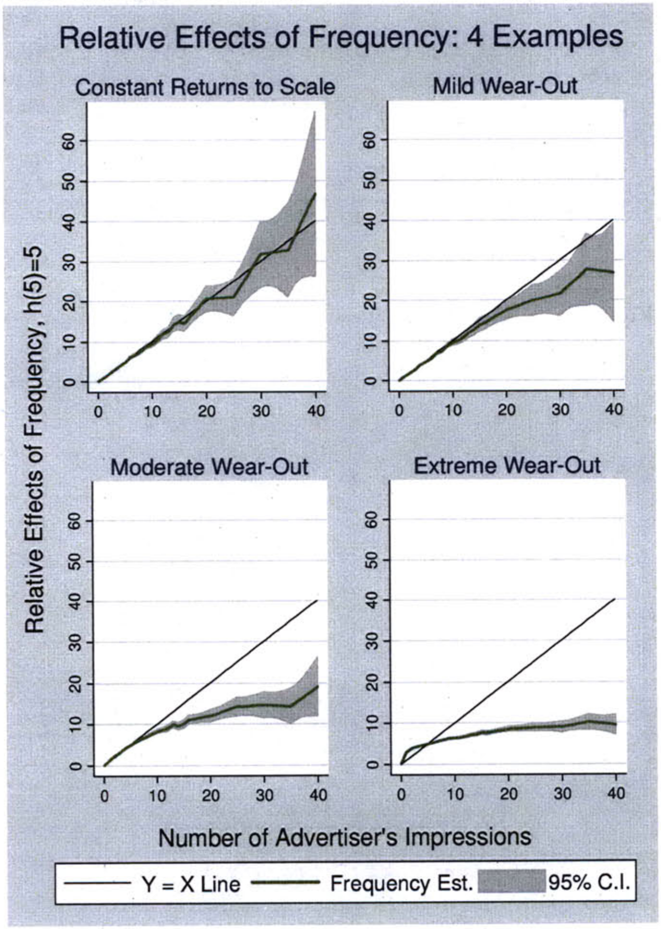


Figure 1-7 - Estimated heterogeneity by browsing type for all 30 campaigns.

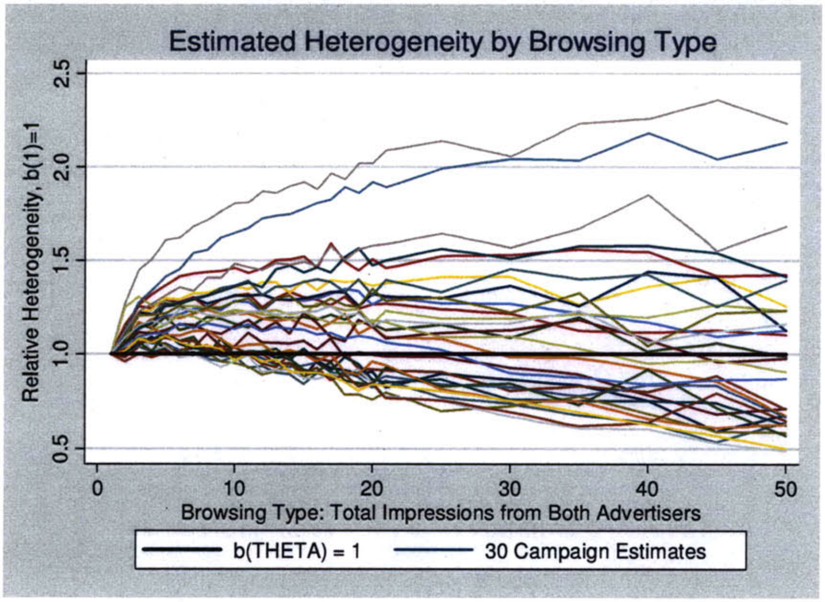


Figure 1-8 - Four examples highlight the range of variation in responsiveness due to browsing type.

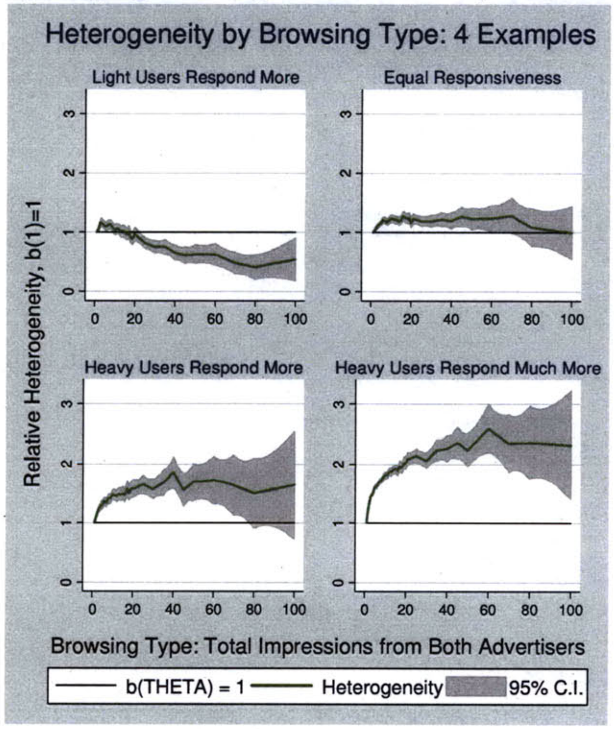


Figure 1-9 - Estimated heterogeneity by browsing type for all 30 campaigns (types 1-10).

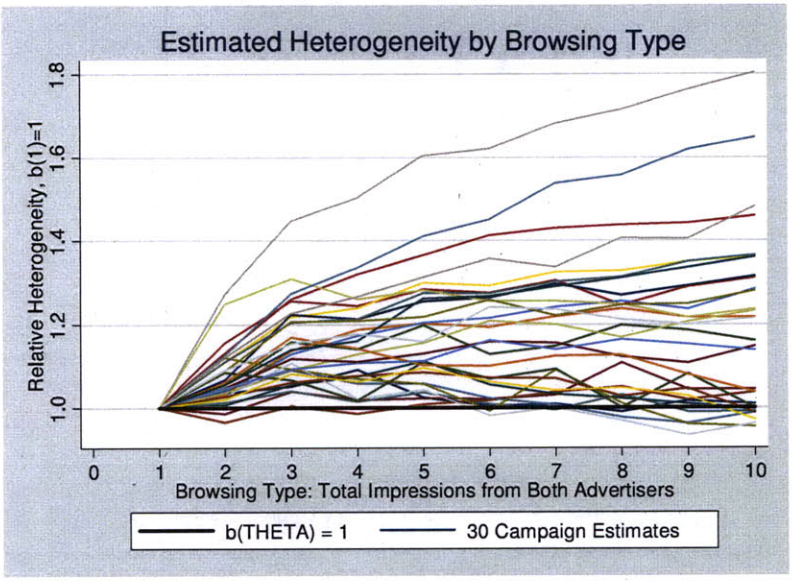


Figure 1-10 - Naive frequency estimates incorrectly measure wear-out.

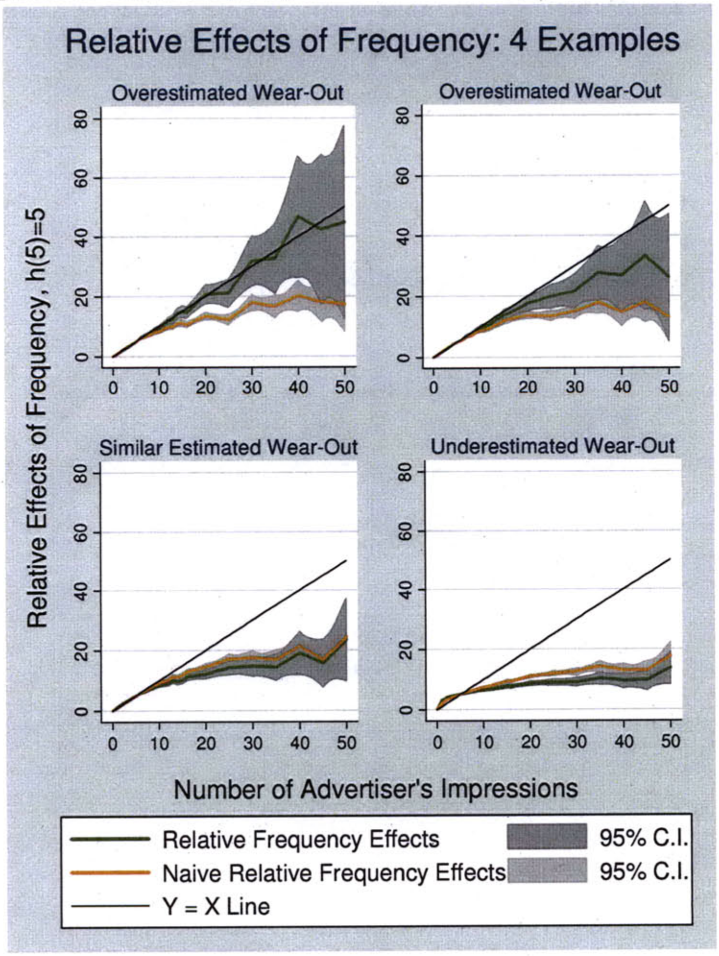


Figure 1-11 - Bias ratio of the naive frequency estimates are large in magnitude.

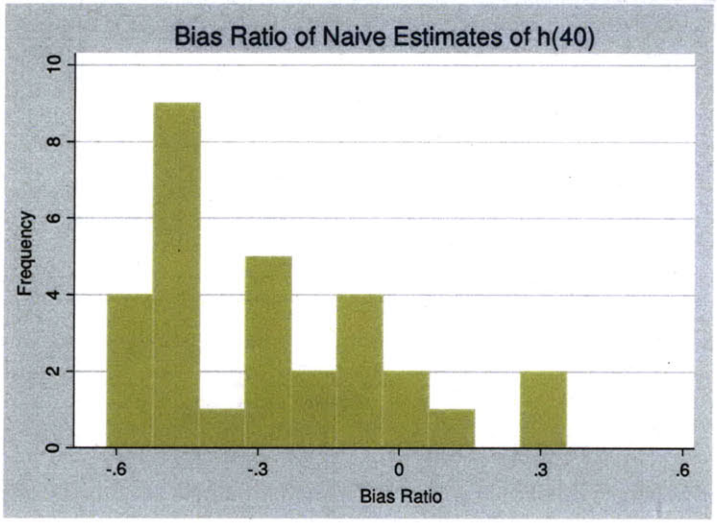
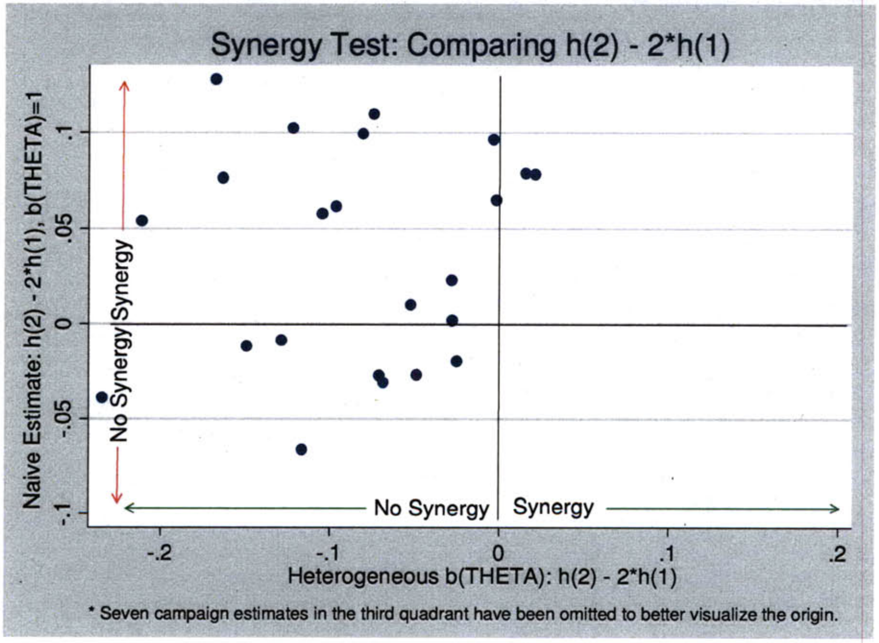


Figure 1-12 - Naive frequency estimates find false ad synergy for many campaigns.



* Seven campaign estimates in the third quadrant have been omitted to better visualize the origin.

Figure 1-13 - Relative effects of frequency on new account sign-ups for 4 campaigns.

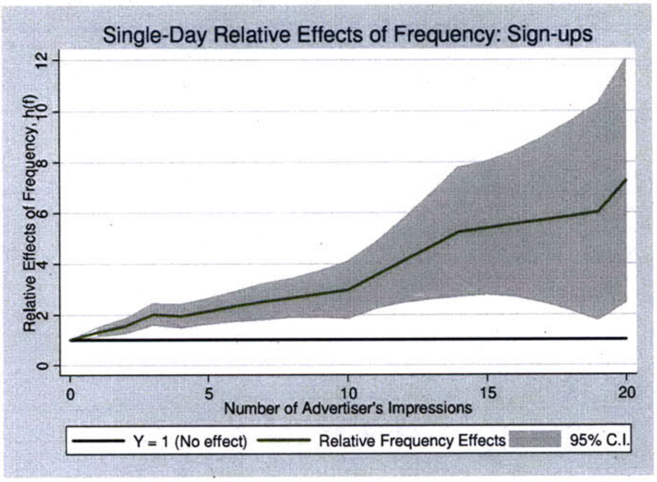


Figure 1-14 - Relative heterogeneity for effects of frequency on new account sign-ups for 4 campaigns.

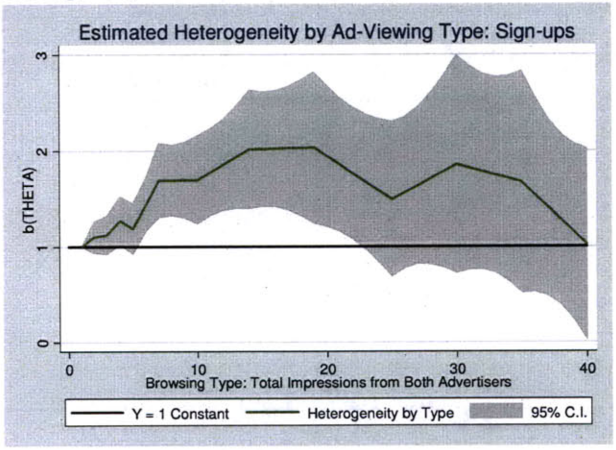
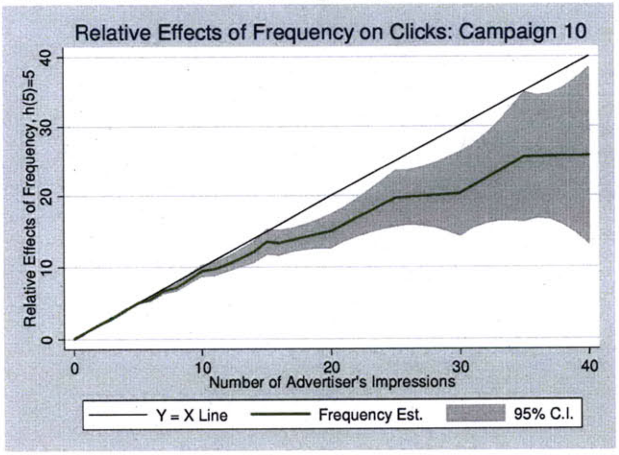


Figure 1-15 - Relative effects of frequency on clicks for campaign 10.



XI. Appendix Figures and Tables

Figure 1-16 - Time profiles of ad delivery for campaigns #27 and #28.

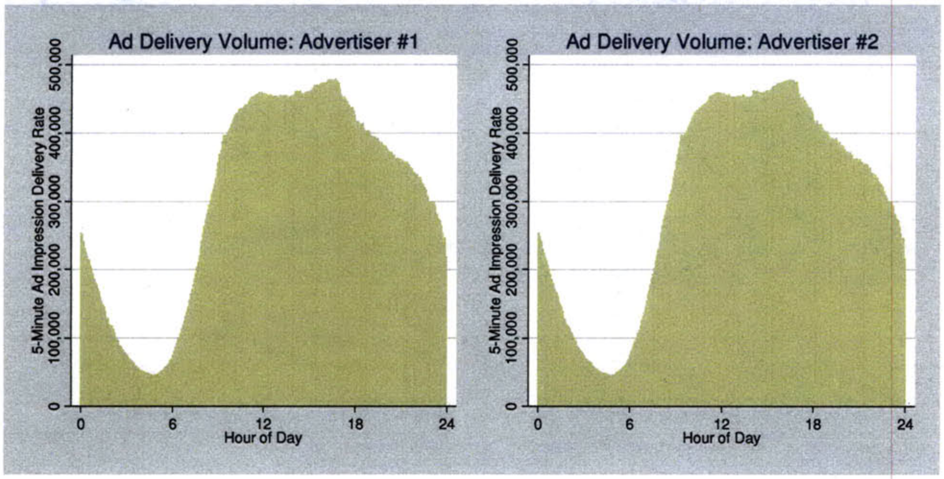


Figure 1-17 - Relative frequency effects for all 30 campaigns.

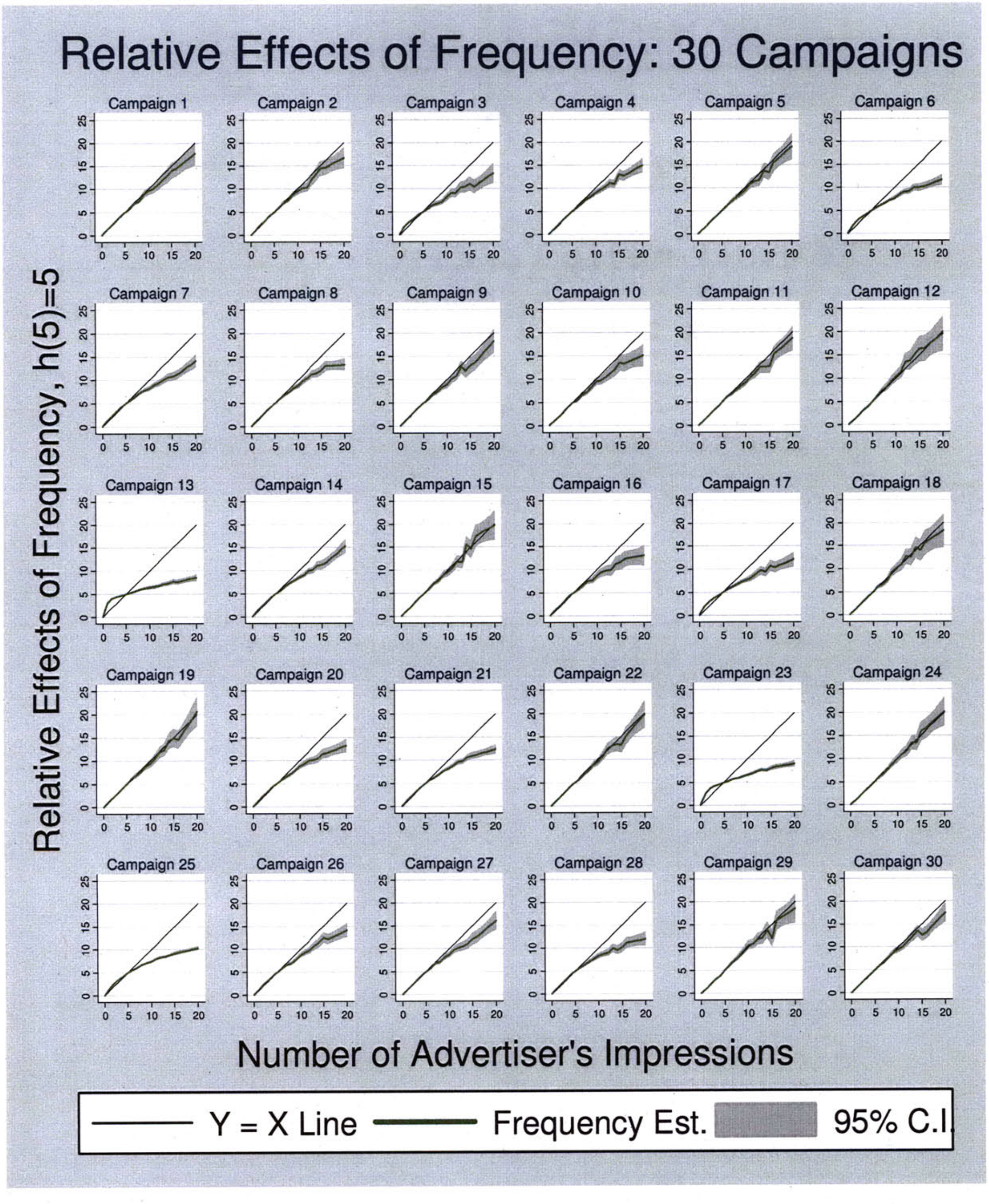


Figure 1-18 - Heterogeneity by browsing type for all 30 campaigns.

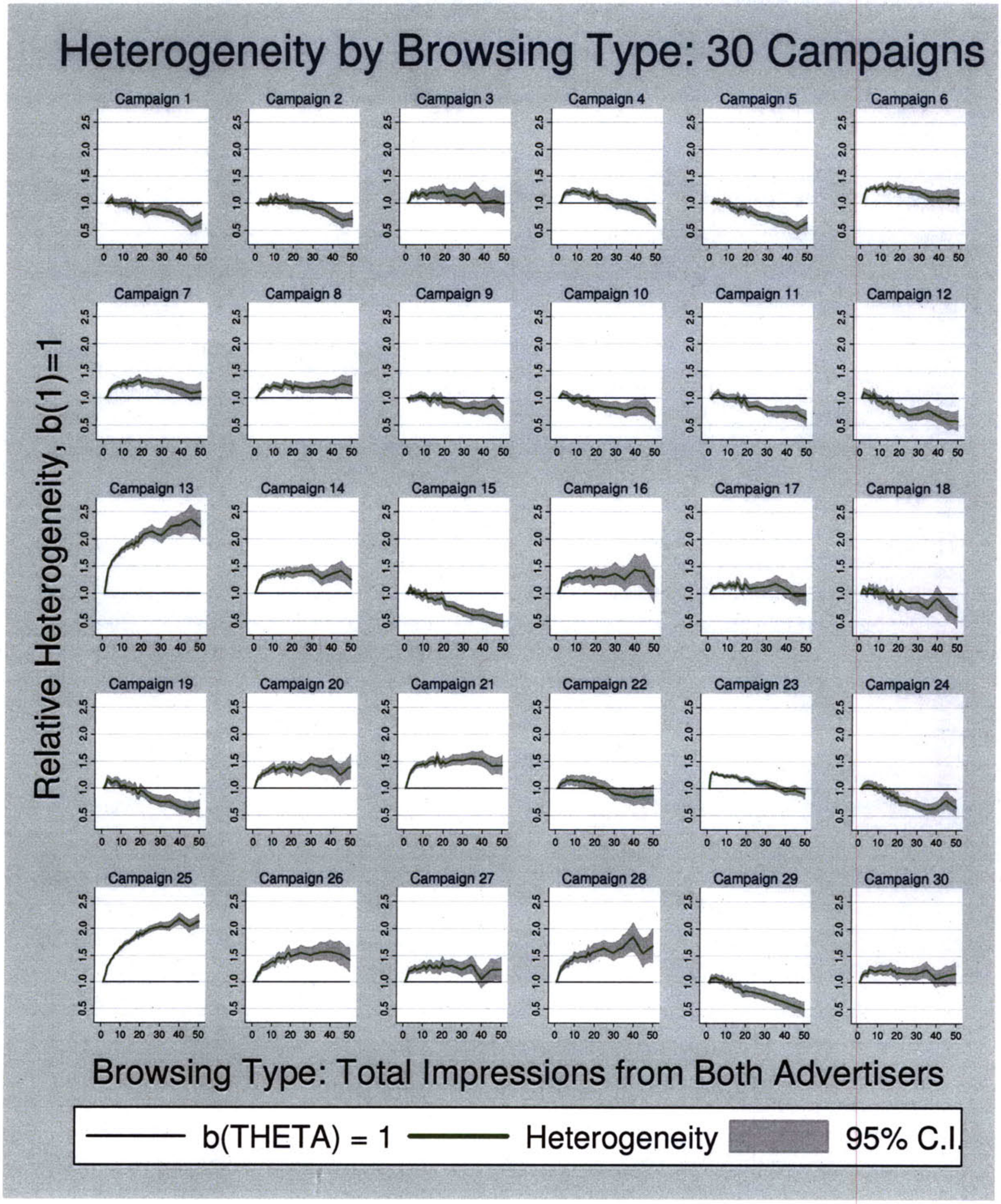


Figure 1-19 - Relative frequency estimates compared with naive estimates for all 30 campaigns.

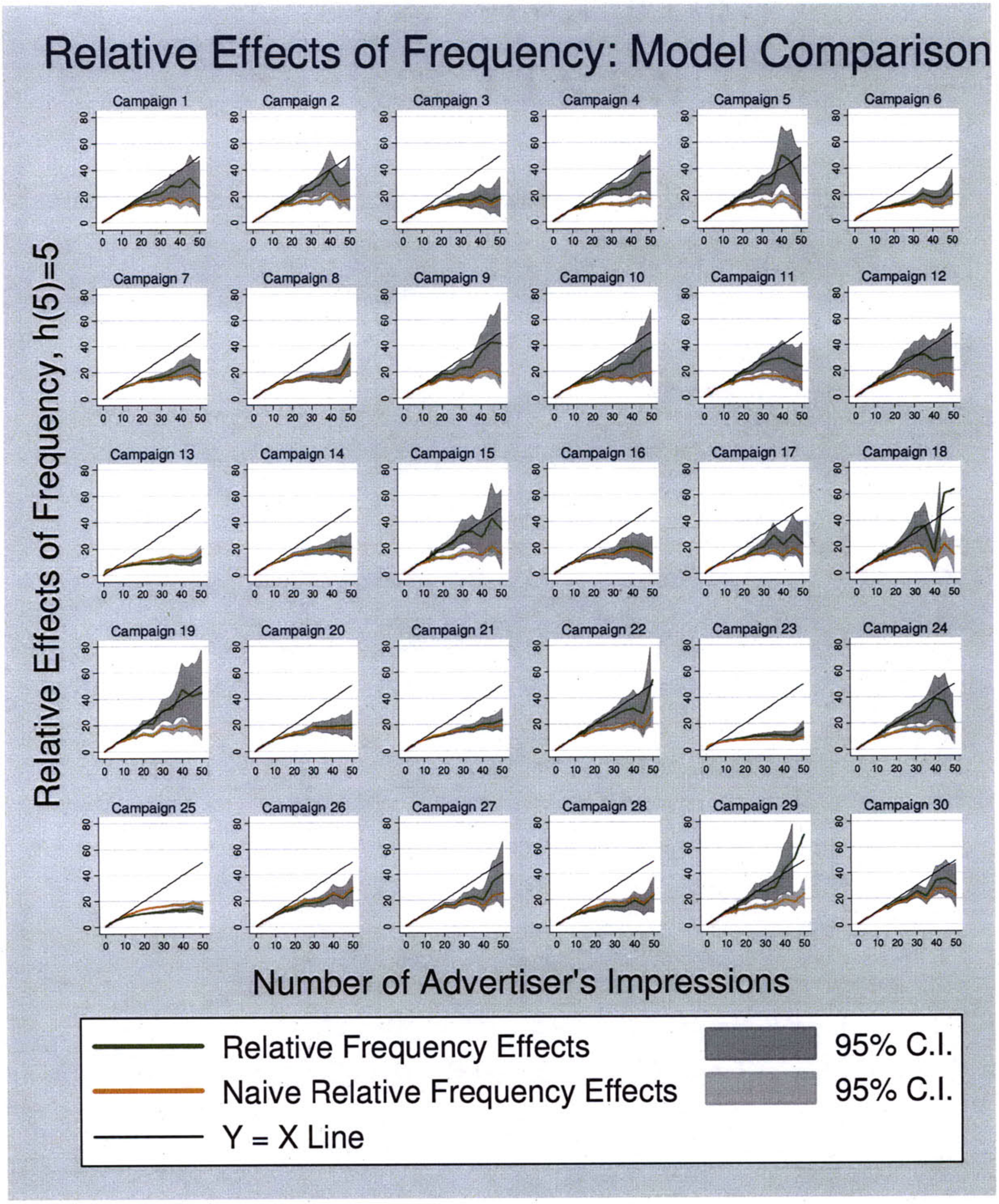


Figure 1-20 - A simple specification test finds the local marginal effects to be in line with the relative frequency estimates for campaign 1.

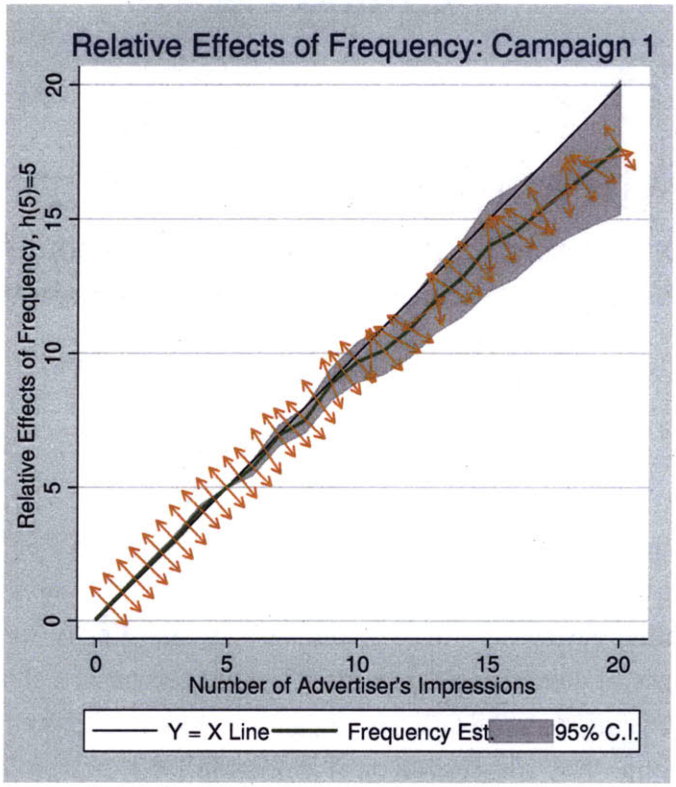
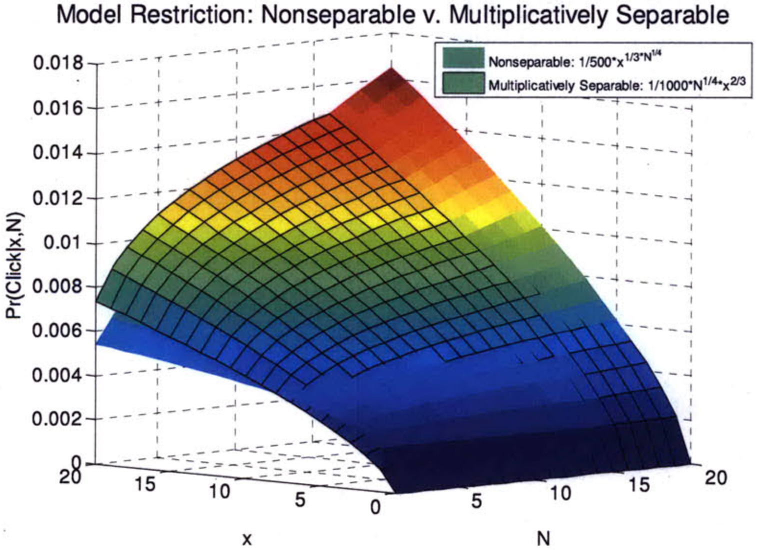


Figure 1-21 - Overlaid Examples of Nonseparable, b∘h(*f*,θ), and Multiplicatively Separable, b(θ)·h(*f*), Models



Chapter 2

Can Online Display Advertising Attract New Customers?

Measuring an Advertiser's New Accounts with a Large-Scale Experiment on Yahoo!

Randall A. Lewis and Taylor A. Schreiner*

Abstract

A large-scale experiment involving 3.7 million treated subjects on Yahoo! tests the ability of online display advertising to attract new customers. The number of new account sign-ups at an online business is tracked and shows a statistically significant impact of one of the two types of advertising campaigns. The ads served as Yahoo! run-of-network succeeded in generating a statistically significant increase in sign-ups of 8-14% relative to the control group. The ads shown on Yahoo! Mail did not produce a statistically significant increase in sign-ups. Despite being derived using millions of subjects, this estimate is quite noisy, with the upper bound of the 95% confidence interval estimate being a 15% increase in new customers. These estimates call into question click-only attribution models, as the number of users that clicked on an ad and converted is less than 30% of the estimated treatment effect.

JEL Classification

C93 - Field Experiments, L86 - Information and Internet Services, M37 - Advertising

Keywords

Online display advertising, advertising effectiveness, field experiment, click attribution

* Lewis: Massachusetts Institute of Technology and Yahoo! Research, randallL@mit.edu and ralewis@yahoo-inc.com. Schreiner: Yahoo! Marketing Insights, tschreiner@yahoo-inc.com. We thank David Reiley for very helpful comments and suggestions. We also thank Meredith Gordon and Sergiy Matusevych for their work on the experiment and the data. Yahoo! Incorporated provided financial and data assistance as well as 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 Clara Lewis, Justin Rao, Glenn Ellison, Jerry Hausman, Stephen Ryan, Sara Ellison, and many participants at the MIT Industrial Organization Workshop.

I. Introduction

Advertising is a frontier in measurement. Tens of billions of dollars are spent each year on a variety of online media such as email, search, display, radio, and video ads intended to reach consumers with information and enticements. In 2009, $22.7 billion were spent advertising online (IAB Internet Advertising Revenue Report 2009) with display ads accounting for 22% or $5.1 billion. In spite of the large scale of these advertising expenditures, little is known about the effectiveness of these dollars. Few studies have gone beyond "the click" to assess the effects of ads on concrete objectives such as sales or new accounts. We address this issue by running a large-scale, randomized controlled experiment tracking ad exposures and outcomes at the individual level.

The experiment randomly assigned each Yahoo! visitor into a treatment or control group. The treatment group was further divided into four subgroups to examine the impact of two different ad placements on the page and two different subsets of Yahoo! properties. The two different ad placements were banner and large rectangular (LREC) ad units. These ads were either served on Yahoo! Mail or served as Yahoo! *run-of-network*, a portfolio of available inventory across many of Yahoo!'s subdomains, including Yahoo! Mail. In total, treatment group users were shown 67 million online display ads for an online business.[1] These ads represented approximately 5% of contemporaneous online advertising expenditure by the company. The ads were targeted at the top 10% of scorers in a standard click-targeting model. Among users eligible to see the ads, 3.7 million Yahoo! visitors were exposed to the ads. Control group users were not eligible to see the ads. The objective was to quantify the impact of the ads on new account sign-ups and identify how to reach the most responsive users.

The banners and LRECs served as Yahoo! run-of-network generated a statistically significant increase in sign-ups of 8-14% relative to the control group. However, the ads shown on Yahoo! Mail did not produce a statistically significant increase in sign-ups. Despite being derived using millions of subjects, this estimate is quite noisy, with the upper bound of the 95% confidence interval estimate being a 15% increase in new customers.

Furthermore, at least 70% of the effect of the ads on new account sign-ups was realized on visitors who saw but did not click on the ads. Accordingly, click-only attribution models

[1] This company also has brick and mortar establishments, but it does most of its business online.

which fail to account for these "view-through" conversions would significantly understate the effects of the ads and should be regarded with caution.

We demonstrate that the technological advances of online advertising can facilitate accurate measurements of the marginal effectiveness of advertising. In particular, randomized controlled experimentation is a viable method for ad campaign evaluation and cost-effective targeting improvements. Using these tools, advertisers can identify subpopulations where the ads had the largest marginal effect and only deliver ads to those who are influenced enough to recoup the advertising expenses through increased profits. Finally, for this advertiser, more than 70% of the incremental sign-ups caused by the advertising come from those who view but do not click ads. As such, click-based attribution models commonly used in the industry would significantly underestimate the effects of this advertising campaign.

In the advertising industry, debates about click attribution attempt to determine each online advertising media's contribution to the overall outcome. Engagement Mapping, created by the Microsoft subsidiary, Atlas (2008), is an example of an online attribution model which assigns different weights to each class of online media events that occurs prior to a conversion. However, the assumptions upon which these models rely have not been scientifically validated and are not universally agreed upon by advertisers and publishers. Publishers seek credit for delivering the advertiser's message to their website visitors while advertisers want to ensure that their advertising works.

This paper contributes to the literature using randomized trials to assess the effectiveness of advertising. Zielske (1959) pioneered the use of field experiments to examine the effects of advertising. He mailed print advertisements to potential customers and measured the impact of the ads over time on consumer recall, or how well an individual could remember an advertiser's name or the content of an ad. Strong (1974) used Zielske's field experiment to calibrate a model of advertising recall which he benchmarked against previous literature. Several advertising experiments run by Eastlack and Rao (1989) explored many different media including outdoor, print, radio, and television from 1975 to 1977 with the Campbell Soup Company. Using distribution warehouse orders as an aggregation of consumer purchases, they found that too much emphasis is placed on media weight, or the amount spent on the advertising, and that more effort should be put into developing high quality creatives.

Subsequent literature, similar to this paper, has examined the effect of advertising directly on consumer purchases. One branch has examined television advertising. Abraham and Lodish (1990) and Lodish, et al. (1995a,b) pioneered experimental measurements of television advertising in the 1980s by exogenously varying cable television subscribers' exposure to ads for consumer packaged goods. Each experiment tracked 3,000 households per market in several markets, matching individual-level ad exposure and purchases using television set-top boxes and supermarket scanners. Their meta-analyses found aggregate evidence of the effectiveness of television advertising across a large number of advertising campaigns. A recent update on this branch of research by Hu, Lodish, and Krieger (2007) confirms the earlier meta-analyses.

Another branch of literature has examined direct-mail advertising's influence on consumer purchases. Simester, Sun, and Tsitsiklis (2006) considered the dynamic effects of mail-order catalog advertising and ran a large-scale field experiment showing that mailing only to immediate purchasers ignores potential long-run profits gained through later purchases from current and future mailings of the catalog. Anderson and Simester (2003) used three field experiments with catalog mailings in which they found that pricing products with endings in $9s increases demand for goods. More recently, Bertrand, et al. (2009) carried out a large-scale randomized direct-mail advertising experiment for short-term loans offered to more than 50,000 South African customers which showed that the advertising's content affected loan applications. Showing fewer example loans, not suggesting a particular use for the loan, or including a photo of an attractive woman increased loan demand by about as much as offering a 25% reduction in the interest rate. However, contrary to psychological predictions, a longer deadline on the offers tends to strongly increase demand.

Online advertising research is most closely related to this work. Danaher and Mullarkey (2003) studied causal contributors to an online display ad's success at capturing attention and improving recall using a lab experiment. Their key finding was that the longer a person is exposed to a web page containing a banner advertisement, the more likely he or she is to remember that banner advertisement. The online advertising medium also provides unique opportunities to track other consumer behaviors such as clicks on display ads to take the viewer to the advertiser's website. Chiou and Tucker (2010) studied the effects of search advertising for pharmaceuticals on health treatment search behaviors. Other recent work has also studied the impact of online advertising on consumer purchases. Lewis and Reiley (2010b) found that the majority of the online advertising effect for a major offline retailer came through their offline storefront. Further, most of this effect could be attributed to the large share of customers who viewed the ad but chose not to click. Lewis and Reiley (2010a) further analyzed the experiment with the nationwide retailer and found that 40% of the impact of the online advertising comes from the 5% of customers who were aged 65 and older.

The remainder of the paper proceeds with an overview of the experiment in section II, a description of the data gathered and used in the experiment in section III, the analysis and results of the experiment in section IV, and concluding remarks in section V.

II. Experiment Overview

We carried out an experiment to assess the effectiveness of online display advertising at attracting new accounts for an online business. The experiment consisted of showing online display advertisements on Yahoo! and then tracking Yahoo! visitors that signed up for a new account at the advertiser's website. We study a company that does most of its business online and expect a majority of the advertising impact to come through the traceable online channel.

The ads were shown on Yahoo!, in some cases at specific subdomains such as news.yahoo.com or mail.yahoo.com, and were primarily large rectangular units (LREC, 300x250 pixels) and banner ads (728x90 pixels). Examples of these ad dimensions and placement for another advertiser (not the advertiser in our experiment) can be seen in Figure 2-1. There we see a Yahoo! Tech page which provides an example of the locations and sizes of both a banner ad across the top and a complementary LREC ad near the bottom of the graphic from the same advertiser.

The experiment used a browser cookie[2] to randomly assign each visitor to Yahoo! to either the treatment group that was eligible to see ads or to the control group which was not. In addition, the ad server was programmed to only show ads to a targeted subpopulation, the top

[2] A browser cookie is a piece of data stored in the user's web browser which helps online services customize a user's browsing experience. Each time a user visits a Yahoo webpage, Yahoo checks to see whether a Yahoo browser cookie identification number is stored in the user's web browser. If the user does not already have a number stored in a browser cookie, Yahoo issues a new identification number and stores it in that user's web browser. Approximately 146 million unique users visit Yahoo each month (comScore 2008), and each has a unique, randomly-assigned number. We use this number to select randomized treatment and control groups. However, the "churn" or deletion of cookies by users is approximately 30% per month for short-lived cookies and 10% for long-lived cookies, which would tend to attenuate our estimated effects. This causes some users to be reshuffled across treatment and control groups during the experiment.

10% scoring visitors[3] in a standard Yahoo! click-targeting model. Then, during the experiment, treatment group visitors were collectively shown 67.4 million LREC and banner ad impressions during their visits to Yahoo! Mail and the other subdomains on Yahoo!. Fifty-two percent of the targeted subpopulation was assigned to the control, leaving the remaining 48% to be treated with ads. This group was broken into four equally-sized treatments: Yahoo! Mail LREC ads, Yahoo! Mail banner ads, Yahoo! run-of-network LREC ads, and Yahoo! run-of-network banner ads. Run-of-network ads are untargeted (by location on Yahoo!) advertising that is shown across the subdomains on the Yahoo! network. This setup, involving an initial random assignment of the entire Yahoo! population with subsequent targeting and ad exposure, is presented in Figure 2-2. Differences between the control and treatment groups represent the impact of the display ads on users' behavior—differences between control group users who "would have seen ads" and treatment group users who "saw ads."

The scale of this campaign is typical for display ad campaigns shown on Yahoo!. Yet, these ads only represented approximately 5% of the advertiser's contemporaneous expenditure on online advertising, and, as such, the experimental ad effects measure the marginal impact of these four targeting alternatives, given contemporaneous exposure to ads purchased with the other 95% of the advertiser's budget.

III. Data

Basic statistics from the experiment are presented in Table 2-1. The 67.4 million ads were shown over six consecutive weeks during January and February 2008 to the four treatment groups, reaching a total of 3.7 million visitors.

Average impressions and click rates differed across media and location. However, the differences in website characteristics and content would lead each of the four combinations of media and location to expose differing (although perhaps overlapping) populations. Thus, direct comparisons cannot be made about the populations that were exposed—only comparisons about what would have happened had one of the other ad campaigns been shown. In particular, we do not know how much of the difference between the Mail and run-of-network campaigns to attribute to the media, location, or ad frequency.

[3] Top 10% of scores for Yahoo! visitors is restricted to "scoreable" users—visitors for whom we had enough information to compute the "custom" click-targeting model, a targeting service offered by Yahoo!. Hence, these ads were eligible to be shown to less than 10% of Yahoo visitors.

Yahoo! Mail visitors who were shown the business's LREC ads saw an average of 8.3 ads, and 0.222% of those users clicked at least once. The sign-up rate for this group was 0.109%. Those who were shown Yahoo! Mail banner ads saw an average of 28.7 ads. The percentage from this group who clicked at least once was 0.715% while 0.102% signed up for an account.

Yahoo! run-of-network visitors who were shown the business's LREC ads saw an average of 8.5 ads, and 0.329% of those users clicked at least once. The sign-up rate for this group was 0.116%. Those who were shown Yahoo! run-of-network banner ads saw an average of 27.3 ads. The percentage from this group that clicked at least once was 0.810% while 0.118% signed up for an account.

These various experiments allow us to gain suggestive insight about characteristics of the ads and their efficacy. A greater frequency would likely encourage more engagement with the ad which could lead to more sign-ups as observed by Lewis (2010). Yet, increasing the average penetration also has a cost, as there are likely many visitors who, by the 20th impression, have already experienced the ad and decided whether they will change their behavior as a result of the ad. This overexposure to the ads is commonly referred to as "ad-fatigue" or "wear-out." This may be a partial explanation for the increased number of clickers on the banner ads which did not result in substantially more sign-ups.

One striking observation is the remarkably small count of users that clicked on the ads that also signed up. For each of the ad campaigns, thousands of individuals clicked on the ads while fewer than 40 of those clickers signed up.[4] Still, clickers account for less than 1% of exposed visitors. To measure the impact of the display ads, we consider both clickers *and* viewers by using a randomized experiment. The experiment provides a proper control group to compare with the treatment groups.

In order to identify the proper control, we use the pre-experiment random assignment of all Yahoo! visitors. Recall, as depicted in Figure 2-2, that while every visitor was assigned to the treatment or control group, only the top 10% of scorers in the targeting model were eligible to be shown the ads. Because we were unable to track a group of control users that would have seen the ads had they been in the treatment group for a direct comparison with treatment users who saw the ads, we resort to other methods. We track the number of Yahoo! visitors who signed up

[4] However, we do not know how many clickers that signed up would have signed up anyway in the absence of the ads. Because clicks are endogenous, it is impossible to construct a control group to measure how many clickers would have signed up had they not been given the opportunity to see or click on the ad.

for an account on the business's website and compare the entire treatment and control group sign-up counts for all users who visited Yahoo! during the six-week campaign, regardless of whether they were shown or even targeted by the ads. Thus, we arrive at a valid, albeit statistically weaker-than-ideal control depicted in Figure 2-3.

IV. Experiment Results

Given a valid experimental control, our analysis is quite straightforward. Recall that we tracked the number of new account sign-ups over six weeks for the treatment and control groups. Since the fraction of the population that signed up for these accounts is small at a mere 0.112% of those treated, our counts can be well-approximated by a Poisson random variable. Thus, in order to compute standard errors for each group's total count, we simply take its square root. Further, since our counts of roughly 20,000 for the treatment and control groups are quite large, the Poisson approximation is actually roughly Normal, allowing for standard Normal-distribution-based inference.[5,6]

The experiment outcomes are presented for reference in Figure 2-4 with 90% confidence intervals and in Table 2-2 with standard errors reported in parentheses as throughout the paper. In total, over the six-week advertising campaign, 20,497 control group individuals (within 52% of the population) and 19,196 treatment group individuals (within 48% of the population) signed up for an account on the business's website.

First, the overall effect of the treatment is the simple difference between the sum of the sign-ups for the four treatments and a scaled (48%/52%) count of the control group's sign-ups. This effect is estimated to be 275.7 (191.5). With a one-sided p-value of 0.075, we reject the null hypothesis that the advertising campaign had no effect at the 10% level with a lower bound of 30.3 sign-ups. With only 82 conversions coming from visitors that clicked on the ads, the share of the ad effect attributable to viewers-only is in excess of 70%.

Second, we decompose the overall effect into the Yahoo! Mail and run-of-network location effects. We find that Yahoo! Mail ads had very little estimable effect of -3.2 (117.6)

[5] The assumption of Poisson may actually be conservative. The nonparametrically-estimated variance of the fifty-two 1% control partitions is 242.2 while the mean, which should be identical under the Poisson model, is 394.2. A heterogeneous Poisson model where, in fact, only a small fraction of the population would be interested in signing up could explain this underdispersion.

[6] This means that to perform hypothesis tests on the difference between our two Poisson random variables, the treatment and control groups' counts, we need not use the Skellam distribution's quantiles for critical values because the difference between two random Normal variables is also Normally-distributed.

sign-ups, while the run-of-network ads had a strong and statistically significant effect of 278.8 (118.8) sign-ups. This strong effect indicates that the run-of-network ads performed much better than the Mail ads. A 90% confidence interval for the Yahoo! run-of-network ads puts the lower bound at 126.6 sign-ups. Further, because only 53 conversions came from clickers, click-only attribution would clearly miss the majority of the ad effect—the large effect from only viewing the display ads.

Third, we decompose the display ad effects into their location and media choice. Both Yahoo! Mail ads performed similarly, posting statistically insignificant effects of 11.9 and -15.1 for LREC and banner ads, respectively. While there would seem to be little estimable effect of the ads, two-sided confidence intervals are large because the estimates' standard errors are 76 sign-ups. While the confidence intervals rule out extreme outcomes, they still include more reasonable treatment effects, perhaps as high as 100 sign-ups for each media choice. Both Yahoo! run-of-network ads performed well, with the LREC ads performing better than the banners with estimates of 173.9 (77.4) and 104.9 (77.0), respectively; however, these estimated counts are not significantly different from one another. Only considering the number of conversions attributable to clickers would grossly understate the effect of the display ads—by 90% and 66% for LRECs and banners, respectively.

In summary, we find that Yahoo! Mail ads failed to produce a large enough effect to statistically differentiate from zero. However, the standard errors of 117.6 sign-ups do not rule out economically significant numbers of sign-ups. Yahoo! run-of-network ads, on the other hand, succeeded for the business by generating a sizeable number of new sign-ups, totaling 278.8 for the 2.2 million individuals that were exposed to the ads. While this may appear to only be a 2.9% increase relative to the scaled control's baseline of 9,460 for all Yahoo! users, this is not the relevant baseline since it does not correspond only to those who were exposed to the ads. Shown in Table 2-1, only 2,558 conversions came from those who saw run-of-network ads. By looking at the two media separately, we find a 13.9% increase in conversions among the LREC ad viewers and an 8.0% increase among the banner ad viewers.

While the marginal effect of the ads may be somewhat small, the lifetime value of a new customer may be large. An estimate of the value of a customer is necessary to determine the cost-effectiveness of the campaign. While precise details about the advertiser's finances cannot be included without compromising anonymity, a back-of-the-envelope calculation suggests that the average value per year of a customer to the advertiser is roughly 25% of the cost per new sign-up. This estimate ignores spillovers such as subsequent referrals and other long-run effects. As such, recouping this advertising investment to acquire new customers may only take four years. However, utilizing the knowledge gained by this experiment could halve that figure for future campaigns. Further, recognizing that run-of-network ads may be as much as 60% Mail ads, targeting other subdomains on which run-of-network ads are shown could substantially improve the cost effectiveness of online advertising with Yahoo! for this online business. This potential for improving campaign performance by specifically targeting users on other Yahoo! subdomains highlights the difficulty in finding and reaching the most responsive population segment while illustrating the valuable returns to learning via experimentation.

V. Conclusion

We demonstrate the limits and potential of online display advertising at generating new customers. Targeting, location, media, creative, frequency, and many other factors can influence a particular campaign's effectiveness. For the configuration of the advertisements in this experiment, we find that the increase in new customer sign-ups attributable to online display advertising for this online business is 11% for users shown Yahoo! run-of-network advertisements. Because the effects of the Yahoo! Mail ads are roughly bounded by 15% and since run-of-network ads include some Mail ads, more effective targeting and advertising product choice has the potential to deliver even greater efficacy of the ads to this particular business. Even higher frequency of Mail ads may compensate for their weaker performance. Future research will focus on identifying and understanding underlying causes of the difference in performance. For now, these campaigns show that an advertiser's incremental online display ad campaign can reasonably generate 8%-14% increases in new customer accounts.

We also demonstrate that attribution models not based on experimental data may misattribute conversions. In particular, standard click-only models can overstate the effect of advertising by glossing over users that clicked that would have converted in the absence of the advertising anyway. Our experiment circumvents this first issue while solving a second: click-based attribution understates the effect of advertising by ignoring the view-only effects of the ads on the typical 99%+ of users that see but do not click on the ad. We find that this effect on viewers is at least 70% for this advertiser's campaign on Yahoo!

We find that the impact of these ads on the 3.7 million viewers reached was realized on as few as 300 individuals. Because the advertising only influenced 0.01% of those exposed to the ads, potentially four orders of magnitude remain to improve the targeting in order to only deliver the ads to those who will be influenced. Rather than targeting the "core demographic" of customers who purchase the most, advertisers can use randomized experiments to obtain precise measurements and identify the different ad products that are most cost-effective and reach the subpopulations that respond the most.

VI. References

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

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

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

Anderson, Eric T. and Duncan I. Simester. 2003. "Effects of $9 Price Endings on Retail Sales: Evidence from Field Experiments," *Quantitative Marketing and Economics*, 1(1): 93-110.

Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir, and J. Zinman, 2009. "What's Psychology Worth? A Field Experiment in the Consumer Credit Market," Yale University Economic Growth Center Working Paper No. 918. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=770389

Chiou, Lesley and Catherine Tucker. 2010. "How does Pharmaceutical Advertising affect Consumer Search? Working paper: https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=IIOC2010&amp;paper_id=514

comScore. 2008. "comScore Media Metrix Ranks Top 50 U.S. Web Properties For December 2008." Press Release: http://www.comscore.com/press/release.asp?press=2685.

Danaher, P.J. and G.W. Mullarkey. 2003. "Factors Affecting Online Advertising Recall:

A Study of Students," *Journal of Advertising Research*, 43(3): 252-267.

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

Eastlack, J. O. and A. G. Rao. 1989. "Advertising Experiments at the Campbell Soup Company," *Marketing Science*, 8(1): 57-71.

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

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

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

Lewis, Randall A. and David H. Reiley. 2010b. "Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via a Controlled Experiment on Yahoo!" Yahoo! Research working paper.

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

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

Microsoft, 2008. "Engagement Mapping: A new measurement standard is emerging for advertisers," atlas Thought Papers, Microsoft Advertiser and Publisher Solutions.

PriceWaterhouseCoopers, L.L.C. 2010. "IAB Internet Advertising Revenue Report 2009." http://www.iab.net/insights_research/1357

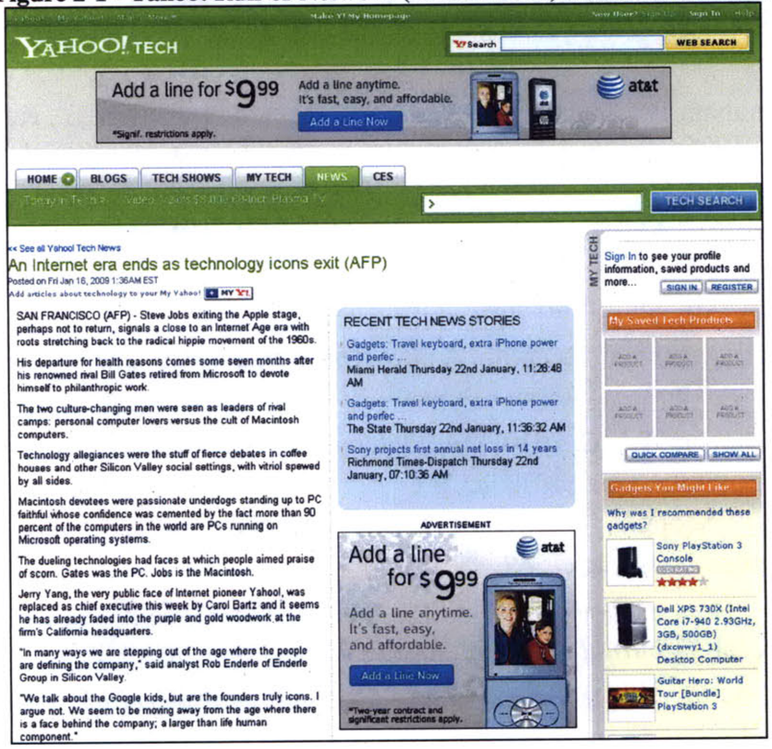
Simester, Duncan I., Peng Sun, and John N. Tsitsiklis. 2006. "Dynamic Catalog Mailing Policies," *Management Science*, 52(5): 683-696.

Strong, Edward C., 1974. "The Use of Field Experimental Observations in Estimating Advertising Recall," *Journal of Marketing Research*, 11(4): 369-378.

Zielske, Hubert A. 1959. "The Remembering and Forgetting of Advertising," *Journal of Marketing*, 23(1): 239-43.

VII. Figures and Tables

Figure 2-1 - Yahoo! Run-of-Network (Yahoo! Tech) Banner and LREC



Add a line for $9.99 Add a line anytime. It's fast, easy, and affordable. Add a Line Now

*Signif. restrictions apply.

<< See all Yahoo! Tech News

An Internet era ends as technology icons exit (AFP)

Posted on Fri Jan 16, 2009 1:36AM EST

Add articles about technology to your My Yahoo!

SAN FRANCISCO (AFP) - Steve Jobs exiting the Apple stage, perhaps not to return, signals a close to an Internet Age era with roots stretching back to the radical hippie movement of the 1960s.

His departure for health reasons comes some seven months after his renowned rival Bill Gates retired from Microsoft to devote himself to philanthropic work.

The two culture-changing men were seen as leaders of rival camps: personal computer lovers versus the cult of Macintosh computers.

Technology allegiances were the stuff of fierce debates in coffee houses and other Silicon Valley social settings, with vitriol spewed by all sides.

Macintosh devotees were passionate underdogs standing up to PC faithful whose confidence was cemented by the fact more than 90 percent of the computers in the world are PCs running on Microsoft operating systems.

The dueling technologies had faces at which people aimed praise of scorn. Gates was the PC. Jobs is the Macintosh.

Jerry Yang, the very public face of Internet pioneer Yahoo!, was replaced as chief executive this week by Carol Bartz and it seems he has already faded into the purple and gold woodwork at the firm's California headquarters.

"In many ways we are stepping out of the age where the people are defining the company," said analyst Rob Enderle of Enderle Group in Silicon Valley.

"We talk about the Google kids, but are the founders truly icons. I argue not. We seem to be moving away from the age where there is a face behind the company; a larger than life human component."

RECENT TECH NEWS STORIES

- Gadgets: Travel keyboard, extra iPhone power and perfec ...
  Miami Herald Thursday 22nd January, 11:28:48 AM
- Gadgets: Travel keyboard, extra iPhone power and perfec ...
  The State Thursday 22nd January, 11:36:32 AM
- Sony projects first annual net loss in 14 years
  Richmond Times-Dispatch Thursday 22nd January, 07:10:36 AM

ADVERTISEMENT

Add a line for $9.99

Add a line anytime. It's fast, easy, and affordable.

Add a Line Now

*Two-year contract and significant restrictions apply.

Sign In to see your profile information, saved products and more...

Why was I recommended these gadgets?

- Sony PlayStation 3 Console
- Dell XPS 730X (Intel Core i7-940 2.93GHz, 3GB, 500GB) (dxcwwy1_1) Desktop Computer
- Guitar Hero: World Tour [Bundle] PlayStation 3

Figure 2-2 - Treatment and Control: Entire Yahoo! Population's Browser Cookie Randomization

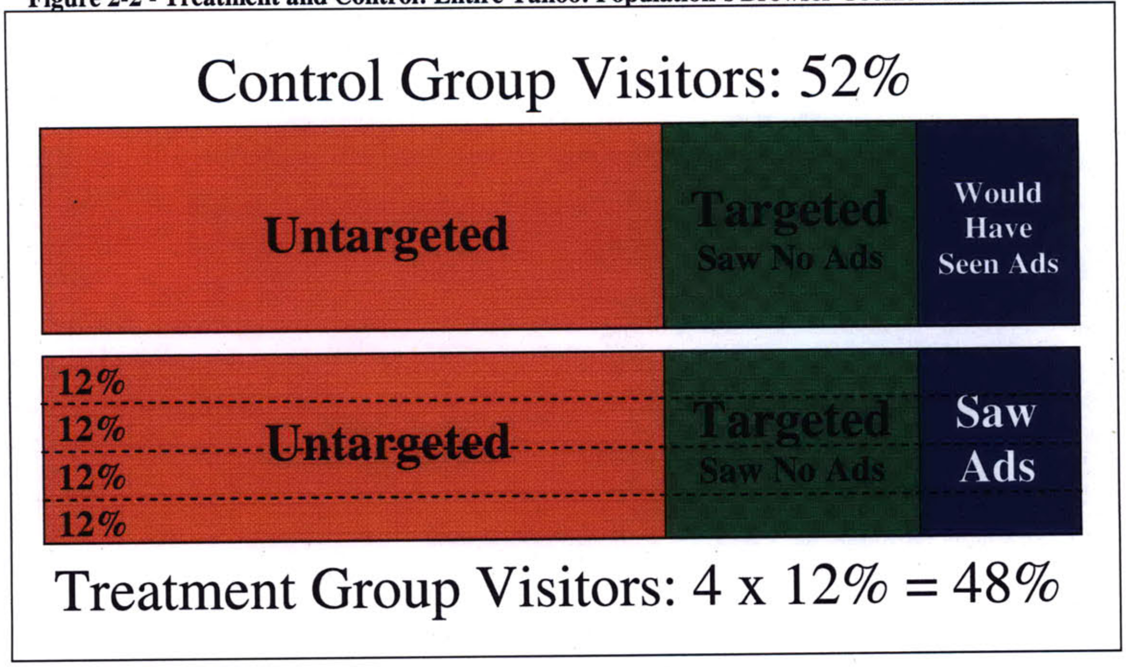


Table 2-1 - Treated Individuals Data Summary

Location	Yahoo! Mail	Yahoo! Mail	Yahoo! Run-of-Network	Yahoo! Run-of-Network
Media	LREC	Banner	LREC	Banner
Number of Impressions	6,613,761	21,506,041	9,169,441	30,099,923
Number of Viewers	794,332	748,730	1,080,250	1,101,638
Avg. Impressions	8.3	28.7	8.5	27.3
Number of Clicks	1,965	6,171	3,992	10,560
Number of Sign-ups	867	762	1,254	1,304
Num. of Clicker Sign-ups	13	16	17	36
Click-Through Rate (CTR) = Num. Clicks / Num. Impressions	0.030%	0.029%	0.044%	0.035%
Clickers/Viewers	0.222%	0.715%	0.329%	0.810%
Sign-up Rate	0.109%	0.102%	0.116%	0.118%

Figure 2-3 - Treatment and Control: All New Account Sign-ups on Online Business's Website

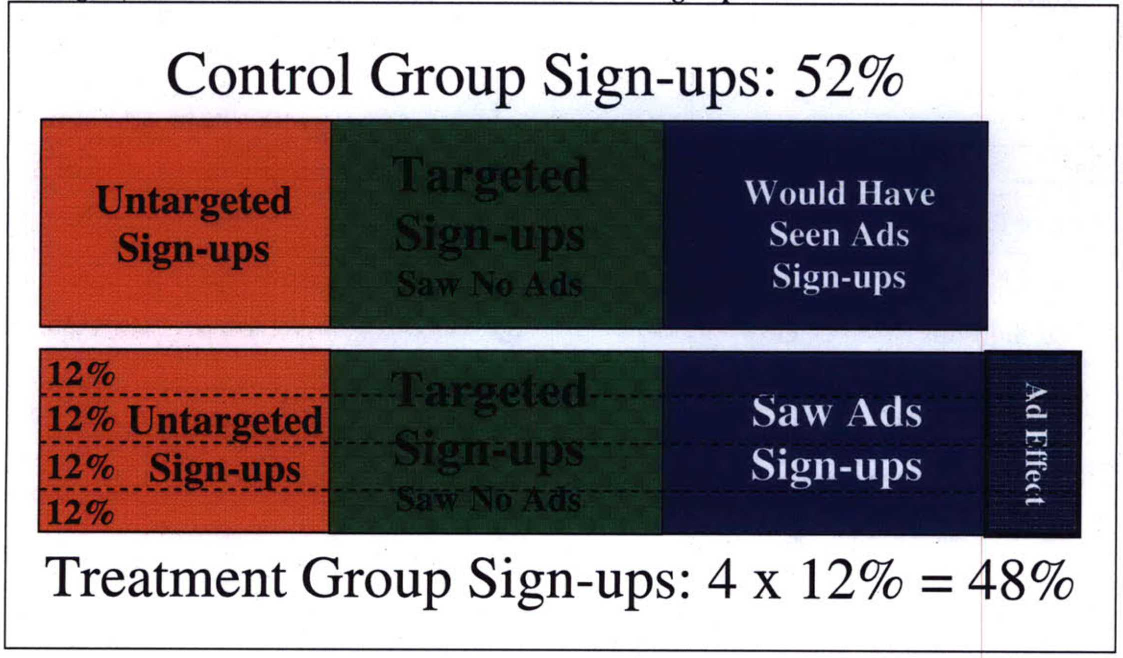

Control Group Sign-ups: 52%
Untargeted Sign-ups
Targeted Sign-ups Saw No Ads
Would Have Seen Ads Sign-ups
12%
12% Untargeted
12% Sign-ups
12%
Targeted Sign-ups Saw No Ads
Saw Ads Sign-ups
Ad Effect
Treatment Group Sign-ups: 4 x 12% = 48%

Figure 2-4 - All Four Treatments versus Scaled Control

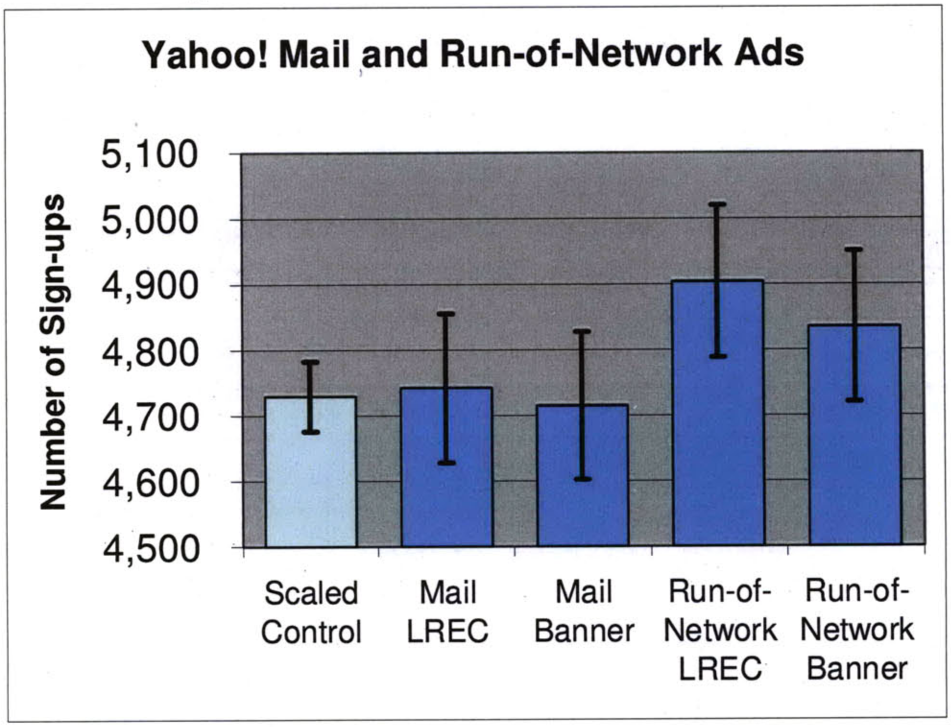


Table 2-2 - All Sign-Up Results

	Control	Treatment			
Location		Yahoo! Mail		Yahoo! Run-of-Network	
Media		LREC	Banner	LREC	Banner
Fraction of Population	52%	12%	12%	12%	12%
Number of New Sign-ups	20,497	4,742	4,715	4,904	4,835
Expected Sign-ups (from Control)	-	4,730	4,730	4,730	4,730
Num. of Clicker Sign-ups	-	13	16	**17**	**36**
Total Ad Effect (Ads)	-	**275.7*** **(191.5)**			
Ad Effect (Location)	-	-3.2 (117.6)		**278.8***** **(123.4)**	
Ad Effect (Location x Media)	-	11.9 (76.4)	-15.1 (76.2)	**173.9**** **(77.4)**	**104.9*** **(77.0)**

Standard errors denoted by parentheses.
*, **, and *** Denotes significance at the one-sided 10%, 5%, and 1% levels, respectively.

Chapter 3

Advertising Especially Influences Older Users

A Yahoo! Experiment Measuring Retail Sales

Randall A. Lewis* and David H. Reiley†

Abstract

Does advertising affect sales in a measurable way? New technologies for tracking both sales and advertising at the individual level are used to investigate the effectiveness of brand advertising for a nationwide retailer. A controlled experiment on 1,577,256 existing customers measures the *causal* effect of advertising on actual purchases, overcoming the major hurdles regarding attribution typically encountered in advertising effectiveness research by exogenously varying exposure to the ads. Online brand advertising generated a statistically and economically significant effect on in-store sales for this retailer. The design of the experiment permits a demographic breakdown of the advertising's heterogeneous effects. Somewhat surprisingly, the effects are especially large for the elderly. Customers aged 65 and older, comprising only 5% of the experimental subjects, exhibited a 20% average increase in sales due to the advertising campaign, which represents 40% of the total effect among all age groups.

Major Classification: Social Sciences

Minor Classification: Economic Sciences

* Lewis: Massachusetts Institute of Technology, randallL@mit.edu, Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, E52-391, Cambridge, MA 02142; Yahoo! Research, ralewis@yahoo-inc.com, 4401 Great America Parkway, 2GA-2616, Santa Clara, CA 95054.

† Reiley: Yahoo! Research, reiley@yahoo-inc.com, 4401 Great America Parkway, 2GA-2614, Santa Clara, CA 95054. We thank Meredith Gordon, Sergei Matusevych, and especially Taylor Schreiner for work on the experiment and data. Yahoo! Incorporated provided financial and data assistance, as well as guaranteeing academic independence prior to our analysis, so that the results could be published no matter how they turned out. We acknowledge the helpful comments of Glenn Ellison, Jerry Hausman, Stephen Ryan, and Duncan Watts.

I. Introduction

Retailing pioneer John Wanamaker (1838-1922) famously remarked, "Half the money I spend on advertising is wasted; the trouble is I don't know which half." Measuring the impact of advertising on sales has remained a difficult problem for more than a century (Bagwell 2008). The econometric problems of endogeneity and omitted-variable bias make it very difficult to establish causality, rather than mere correlation, from observational data. In the present paper, we overcome this problem with a large-scale controlled experiment for an online advertising campaign and decompose the estimated effects of the advertising by age and gender. We find the surprising result that this advertising has its largest impact on older people.

As an example of the opportunity to draw mistaken conclusions from non-experimental data, we consider a recent state-of-the-art study by marketing practitioners (Abraham 2008). We quote this study, which estimates large positive effects of advertising on sales, to describe its methodology, "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."

With observational data, this straightforward technique can give spurious results. The population who sees a particular ad may be very different from the population who does not see it. For example, those people who see an eTrade ad when searching for "online brokerage" are very different from those who do not see that ad, even in the absence of advertising. Almost certainly, 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." So the observed difference in sales may not represent a causal effect of ads at all (Lewis and Reiley 2010).

Many studies of advertising, both in industry and in academia, suffer from similar problems of establishing correlation rather than causality. In general, advertisers do not systematically vary their levels of advertising to measure the effects on sales. Advertisers often change their levels of advertising over time, running discrete "campaigns" during different calendar periods, but this variation does not identify causal effects because other relevant variables also change. For example, if a retailer both advertises more and sells 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. Similarly, observational academic studies providing regressions of sales data on advertising data often produce positive relationships, but they might be due to reverse causality, when advertisers set advertising budgets as a fraction of sales (Berndt 1991; Dorfman and Steiner 1954; Schmalensee 1972).

A time-honored scientific method for establishing causal effects is to run a controlled experiment, with treatment and control groups. In the eTrade search-advertising example, we would ideally like to experiment on the population who searches for "online brokerage," showing the eTrade ad only to a treatment group, but not to a control group. The idea of replacing observational studies with field experiments has recently gained importance in economics and the social sciences (Levitt and List 2008).

Measuring the effects of brand advertising presents particular challenges, because the effects are likely to be quite diffuse. A given advertising campaign may be a very small factor in a given customer's purchase decision, and the results may not be as immediate as in direct-response advertising. A large sample size should enable detection of an advertising signal from the sales noise, but historically it has been nearly impossible to measure both sales and advertising for each individual.

Past attempts to use experiments to measure the effects of brand advertising on sales have found mixed results because of the low signal-to-noise ratio. In the 1970s, a series of 19 advertising experiments for Campbell's Soup measured aggregate sales separately for up to 31 different geographical areas (Eastlack and Rao 1989). In the 1980s and 1990s, IRI's BehaviorScan technology enabled advertisers to deliver differing quantities of television ads and measure the effect on sales for 3,000 households at a time (Abraham and Lodish 1990; Lodish, et al. 1995a,b; Hu, Lodish, and Krieger 2007). These early experiments generally lacked sufficiently large samples to reach statistically conclusive results.

In the present study, we take advantage of recent technological developments in online advertising and customer database management in order to overcome this signal-to-noise problem. Our controlled experiment tracks the purchases of over one million individuals in order to measure the effectiveness of a retailer's nationwide display-advertising campaign on Yahoo!

II. Experimental Design

We executed a field experiment using a sample of a nationwide retailer's existing customers, identified via matching either postal or email addresses between the retailer's customer database and Yahoo!'s user database. This match yielded a sample of 1,577,256 individuals for whom we could track weekly purchases, both online and in the retailer's physical stores. Of these matched users, we randomly assigned 81% to a treatment group who were exposed to an advertising campaign for this retailer on Yahoo! The remaining 19% of users were assigned to the control group, who viewed no ads in this advertising campaign. After the campaign, a third party matched the retailer's sales data with Yahoo's advertising data for each individual in the sample and then anonymized each record to protect customer privacy in the final dataset.

The advertising campaign delivered 32 million advertisements to the treatment group during 14 days of October 2007, reaching approximately 800,000 unique users. These rectangular, graphical advertisements appeared on various Yahoo! websites such as mail.yahoo.com, groups.yahoo.com, and maps.yahoo.com. Fig. 3- shows a similar ad on Yahoo! Shopping. The large, 300-by-250-pixel Capital One graphic ad is an example of a typical "run-of-network" ad of the size and shape used by the retailer in our experiment. These ads are distributed across all pages on the Yahoo network, typically located on the right, near the top of the page. The ads appeared only to treatment-group users when logged in to Yahoo!; none of the retailer's ads were shown to users who were not logged in. No other users saw this advertising campaign's ads, whether logged in or not, so we are confident that control-group is valid for identifying customer behavior in the absence of the ads.

These represent the only display ads shown by this retailer on Yahoo! during this time period. However, Yahoo! ads represent a small fraction of the retailer's overall advertising budget, which include other media such as newspaper and direct mail. Yahoo! advertising turns out to explain a very small fraction of the variance in weekly sales, but because of the randomization, the treatment is uncorrelated with any other influences on shopping behavior.

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

III. Data

Demographic statistics available to us indicate a valid randomization between treatment and control groups. Chi-squared tests generate no significant differences in gender (p=0.234), age (p=0.790), or U.S. state of residence (p=0.232). Nor are there significant differences in Yahoo! usage statistics, including the total number of pages viewed, properties visited, and categories of keywords searched.

Fig. 3-2 displays the distribution users in the experiment by age and gender, as self-reported by users to Yahoo! The age distribution for this retailer is roughly representative of the general working population, with most customers falling in the range of 25 to 60 years of age. Approximately 93% of customers are at least 25 years of age, 53% of users are at least 40, and 6% of users are at least 65. Both genders are well represented in every age group, with women composing 59.6% of the experimental population.

Treatment-group members saw ads from this campaign on 7.0% of the Yahoo! pages they viewed during the campaign period. The probability of seeing this retailer's ad on a given page view depended on several factors, including user characteristics as well as the type of page they visited on Yahoo! On the other 93% of pages visited during the same time period, treatment-group members saw a variety of ads from other advertisers.

The number of ads viewed by each Yahoo! user in this campaign is quite skewed. The majority of users in the treatment group viewed fewer than 100 ads in the two-week campaign. By contrast, 1.0% of the treatment group viewed more than 500 ads, with a maximum number of 6050 ad views for one person in two weeks. Such extreme numbers suggest the possibility that a few of our users were automated browsing programs, or "bots," despite the fact that these user accounts were known to match real people who have previously shopped at the retailer. Indeed, Yahoo! engages in fraud-prevention activities that identified some impressions as not being valid deliveries to humans and did not charge the retailer for those. We were not able to exclude these potentially invalid observations from our dataset, but they appear to represent a very small part of the sample. To the extent that we include treatment-group members who could not actually have perceived the ads, we will have a slight bias towards measuring zero effect of the advertising.

The distribution of customer purchases includes a large number of zeroes: only 4.8% of the customers in our data engage in a transaction during the two weeks of the campaign. The standard deviation of sales across individuals, R$19.49, is much larger than the mean of $1.88. Though there are some very large outliers, over 90% of positive purchase amounts lie between –R$100 and +R$200, where negative purchases represent returns to the retailer. Total purchases occur 14% online versus 86% offline (in-store).

IV. Results

The treatment group purchased R$1.89 per person during the campaign versus R$1.84 per person for the control group, indicating a positive effect of the ads of approximately R$0.05 per person in the treatment group. Since only 64% of the treatment group actually viewed ads, this represents R$0.08 per person treated. This treatment effect is statistically significant at the 10% level (p=0.092, one-sided); below we shall see that the statistical significance is much greater for older customers. Multiplying by the total number of customers treated, our best estimate of the aggregate, contemporaneous increase in sales due to advertising is R$83,000, compared with a cost of R$25,000. Assuming a contribution margin of 40% for this class of retailer and ignoring any long-term effects of the ads on shopping behavior, we estimate the ad campaign's return on investment (ROI) to be well over 30%.

Next, we decompose the treatment effect by age (Fig. 3-3, top panel). This plots the difference between two locally linear regressions, each using an Epanechnikov kernel with a bandwidth of four years. We find substantial heterogeneity: individuals between ages 20 and 40 experienced little or no effect from the advertising, while individuals aged 50 to 80 experience a sizeable positive effect on sales. Baseline rates of offline and online sales (Fig. 3-3, bottom panel, Epanechnikov bandwidth of two years) indicate that older customers buy no more than younger customers. The effect of advertising on sales increases with age. During the 14-day advertising campaign, treated individuals over 40 years old increased their purchases by an average of R$0.15 per person relative to the control group. Customers aged 65+ responded the most, with an average increase in sales of R$0.37.

Presented for comparison are baseline purchase levels, averaged across treatment and control. Offline purchases are nearly five times as large as online purchases. Teenagers outspent all other age groups, but they are a very small fraction of the customer population. Customers aged 40-65 purchase more than customers 20-40, with a slight decline in purchase amounts after age 65.

A simple model would assume that advertising's effects scale proportionally with sales volume. For example, if middle-aged customers purchase 50% more than the elderly, then the impact of advertising on the middle-aged customers should experience an effect that is 50% larger. However, as a percentage of total purchases, the treatment-effect difference between older and younger customers, shown in

Fig. 3-4, is even more pronounced than in dollar amounts. The treatment effect is 35% of total purchases for customers aged 70 to 75, but only about 5% for customers aged 40 to 65.

This figure uses the two panels of Fig. 3-3 to compute treatment effects as a percentage of sales, rather than an absolute dollar difference, with pointwise error bands computed by the Delta Method. The 95% confidence interval gives an upper bound on the treatment difference: no more than 10% of average sales for individuals under age 65. By contrast, individuals aged 70 to 75 have an estimated effect of approximately 35%, three times the upper bound for younger individuals.

We consider two trivial reasons why the advertising might influence the purchases of older customers more than those of younger customers. Do the older customers view more advertisements? The data shows the opposite: conditional on seeing any of the retailer's ads during the 14-day campaign, customers aged 25 to 35 saw nearly 50 ads on average, while customers aged 50+ saw fewer than 35 ads on average (Fig. 3-5, top panel). Are older customers more likely to see any of the retailer's ads? The answer is again no (Fig. 3-5, middle panel). The probability of viewing any ads at all also decreases with age: the 25 to 40 age group were about 5% more likely to see ads than those over 40.

In addition to reacting more in terms of purchases, older customers also turn out to be more likely to react to ads by clicking on them (Fig. 3-5, lower panel). Conditional on seeing at least one ad, the probability of clicking at least one ad increases monotonically until age 45, and is approximately constant for customers over 45 years old. More than 8% of customers over 45 clicked on ads they saw, compared with only 4% of 20-year-olds. Over the entire sample, 7.2% of ad viewers chose to click.

We have seen evidence that the effects of advertising vary by age. We next quantify these treatment effects by age group, using two simple categorizations of "older" versus "younger" customers, providing standard errors in parentheses. Customers at least 40 years old have an average increase in purchases of R$ 0.092 (0.055), six times higher than the estimate of R$ 0.015

(0.051) for those under 40. Using this split, the treatment effect is statistically significant for older customers (p=0.047, one-sided), but statistically insignificant for younger customers (p=0.388, one-sided). Customers at least 65 years old exhibit an average increase due to advertising of R$ 0.373 (0.127), ten times higher than the estimate of R$ 0.037 (0.039) for those under 65. Senior citizens' increased purchases are statistically significantly different both from zero (p=0.002, one-sided) and from the treatment effect of younger customers p=0.012, two-sided), while the estimated effect for younger customers is statistically insignificant (p=0.175, one-sided).

In addition to age, we further decomposed our estimates by gender. While not statistically significantly different (p=0.118, two-sided), the point estimates are suggestive of the advertising being much more effective for women with an estimated effect of R$ -0.018 (0.062) for men and R$ 0.105 (0.047) for women. The treatment difference between women and men is R$ 0.082 (0.111) for those at least 40 years of age, and R$ 0.390 (0.252) for those at least 65. For this particular campaign, women account for most of the total effect of the advertising, with the greatest effect on women around 70 years of age.

We also investigated the effects of the advertising separately for online versus offline sales. Online sales increased by only R$0.007 (0.013) per person in response to the advertising, while offline sales increased by R$0.046 (0.035) per person. The effect on online sales is quite imprecisely estimated, and when we break it down by age we do not see any obvious difference between older and younger customers. Only the offline sales show the striking trend of more effect of advertising on older customers. (See the Appendix for more detailed graphs of the treatment effects separately for each combination of age, gender, and online/offline sales.) This result seems quite surprising when we compare it to the above result on ad-clicking behavior. Older customers who saw ads in this campaign were much more likely to click them than younger customers, but they did not increase their sales much in the online store where the ad click took them. Rather, the ads appear to have induced them to do more research online and more actual shopping in brick-and-mortar stores.

To check robustness of these results, we subsequently (November 2007 and January 2008) carried out two additional online display ad campaigns with the retailer, using the same treatment-control assignment as in the original campaign. These two campaigns were each much smaller and shorter than the one just analyzed: 10 million and 17 million impressions, respectively, over 10 days each, compared with 32 million impressions over 14 days in the first campaign. The advertiser chose the size and timing of each campaign. Because these campaigns were both smaller and later than the original campaign, we might expect their effects to be smaller. While qualitatively similar to the results for the first campaign—advertising has positive effects on sales, more so for women and older customers—the effects are indeed less statistically significant. More details can be found in the Appendix.

V. Conclusion

Our striking result is that this advertising especially influenced older users. We caution that these results are for one particular campaign for one particular retailer, and we do not know to what extent they will generalize. It is entirely possible that other types of advertising campaigns will produce opposite results, with younger customers responding more than older ones. Nevertheless, we find it scientifically meaningful to be able to use our technology to document at least one case where, in contrast to many marketers' belief in the desirability of advertising to younger customers (NAMC 2009), we see clearly that older users respond more to advertising than younger users.

Why might this be the case? We propose four possible explanations, two from economics and two from psychology. From an economic point of view, it might be the case that retired customers have a lower opportunity cost of time than younger customers, so they are more willing to spend time paying attention to display advertisements. Older customers may also have higher wealth available for discretionary spending.

From a psychological point of view, previous research has shown that repetition of a message tends to induce beliefs more strongly in older individuals than in younger ones (Law, Hawkins, and Craik 1998). As this campaign did involve dozens of repetitions of brand-advertising messages to individual customers, this could be a cause of our age-based differences. Second, previous research has shown older individuals to process information less quickly than younger individuals, and particularly have lower recall of time-compressed television ads (Stephens 1982). To the extent that older customers respond relatively better than younger users when allowed to process information at their own pace, online display ads may produce better effects among older users.

In addition to raising interesting scientific questions, our large-scale, controlled experiment also has practical implications. We close with a back-of-the-envelope calculation of costs and benefits to the retailer of advertising to the different age segments of the population. Each advertising impression cost the retailer approximately R$ 0.001 on average. In the youngest half of our sample (i.e., under 40 years of age), the average user saw 45 impressions, so the cost of advertising to those customers was about R$0.045 per person, by comparison with a revenue increase of only R$0.015. Thus, advertising to the younger customers looks like an unprofitable investment, at least in the short run, although we caution that our lack of statistical precision does not allow us to rule out profitability. The older half of the sample, however, was a much better investment, viewing 35 advertising impressions per person, for a cost to the advertiser of R$0.035 per customer. This cost compares quite favorably with a revenue increase of R$0.092 per customer 40 and older, and R$0.372 per customer 65 and older, even when we consider that the retailer's marginal profit is only about 50% of its marginal revenue. We thus find an extremely large rate of return to advertising to the older customers in the experiment. These numbers suggest a return on investment of over 30% for customers 40 and over, compared with nearly 1,000% for customers 65 and over. Large-scale, controlled experiments like this one therefore show promise for allowing advertisers to eliminate the "wasted half" of advertising theorized by John Wanamaker over a hundred years ago.

VI. References

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

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

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

Berndt, E. R. 1991. *The Practice of Econometrics: Classic and Contemporary*. Reading, Massachusetts: Addison-Wesley, chp. 8.

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

Eastlack, J. O. and A. G. Rao. 1989. "Advertising Experiments at the Campbell Soup Company," *Marketing Science*, 8(1): 57-71.

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

Law, S., S. A. Hawkins, and F. I. M. Craik. 1998. "Repetition-Induced Belief in the Elderly: Rehabilitating Age-Related Memory Deficits." *Journal of Consumer Research*, 25: 91-107.

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

Lewis, Randall A. and David H. Reiley. 2010b. "Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via a Controlled Experiment on Yahoo!" Yahoo! Research working paper.

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

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

NAMC Newswire Releases. 2009. "Brand Young and You Brand for Life; Why Cinema Advertising is an Effective Tool for Marketers," April 17, 2009.

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

Stephens, N. 1982. "The Effectiveness of Time-Compressed Television Advertisements with Older Adults." *Journal of Advertising,* 11(4): 48-55.

VII. Appendix

As noted in the body of the paper, we delivered a total of three campaigns for this retailer to the treatment group over a period of several months. The first campaign was the largest and, hence, provides the cleanest opportunity to estimate a statistically significant effect of the ads, so we concentrate on that campaign in this paper. The other two campaigns provide an opportunity for robustness checks; however, we expected them to have weaker effects due to delivering a much smaller number of impressions. Also, there was no re-randomization between campaigns, so to the extent that advertising may have persistent effects, the effects of Campaigns 2 and 3 will not be separately identified from the unknown long-run effects of Campaign 1. Table 3-1 summarizes the delivery of ad impressions in the three different campaigns.

The sales data include separate measures of offline and online purchases for each individual each week. Sales amounts include all purchases that the retailer could link to each individual customer in the database, using such information as the name on a shopper's credit-card during checkout at the store. To the extent that these customers sometimes make purchases that cannot be tracked by the retailer (for example, using cash and not identifying themselves), our estimate may underestimate the total effect of advertising on sales. However, the retailer believes that it accurately tracks at least 90% of purchases for these customers.

The text reports summary statistics that support valid randomization, especially along the dimensions of gender and age. Here, we investigate this issue in additional detail. In Fig. 3-6, we plot the fraction of the treatment and control groups that were female by age and computed the difference. Neither plot suggests any anomalous treatment-control differences. The left panel shows the fraction of the sample at each age group that was female for the control group. For customers around ages 25 and 55, there are disproportionately more women than men relative to other age groups. The right panel shows the difference between the treatment and control groups regarding the fraction of the sample that is female for each age group. The lack of statistical difference from zero indicates a valid randomization with respect to gender.

Campaigns 2 and 3, which were shown to the treatment and control groups following the campaign, corroborate both the effect of the ads on sales and with respect to age. In terms of the average effect of the ads across all customers (Table 3-2 and Table 3-3), all three campaigns exhibited effects of the ads on sales of approximately 3% on both online and offline channels, with roughly 80-90% of the effect of the ads coming through the offline channel, in line with the offline sales volume of 84% of total sales for the control group. Thus, large retailers which do most of their business offline can reap benefits both online and offline from advertising online.

Following the discovery that the advertising seemed to affect women more and the offline channel more, we examined the heterogeneous differences in purchasing behavior for the three weeks prior to the campaign (Fig. 3-7) and for each of the three campaigns separately (Fig. 3-8, and Fig. 3-9, and Fig. 3-10).

We now examine the robustness of the result that the elderly purchase more in response to the ads for campaigns 2 and 3. At first glance the pre-experiment sales results (Fig. 3-7) appear to partially explain the main results of the paper as the elderly in the treatment group have slightly higher sales than the control group. However, we would like to highlight the fact that less than 5% of the customers purchase in any given week and there are few repeat purchasers. In addition, while the treatment group appears to purchase more than the control for the elderly around age 80, the location of this differential appears to occur at slightly different ages than it does during campaign 1. Thus, we conclude that the statistical variation prior to the experiment does not explain our results which rely on purchases by different people during subsequent weeks.

An examination of the variation in the time dimension allows us to further test the robustness of the results. We compute an experimental difference-in-differences (DID) estimate by subtracting the pre-experiment average weekly sales from the average weekly sales during the campaign for each individual and then comparing these averages across the treatment and control groups. In our experimental DID, we find very similar results to our estimates presented in the main text (Fig. 3-11). We have relegated these results to the Appendix to simplify the exposition by avoiding descriptions of the pre-experimental sales.

Upon completing the in-depth decomposition of the results for all three campaigns, we discovered several other marginally significant regions among the estimates for Campaigns 2 and 3. However, we hesitate to rush to any conclusions, due to the risk of committing multiple type I errors. We specifically demand a greater level of significance from our primary results to avoid any spurious conclusions arising from multiple-hypothesis testing problems.

VIII. Figures

Fig. 3-1. Example of a typical graphical advertisement on Yahoo!

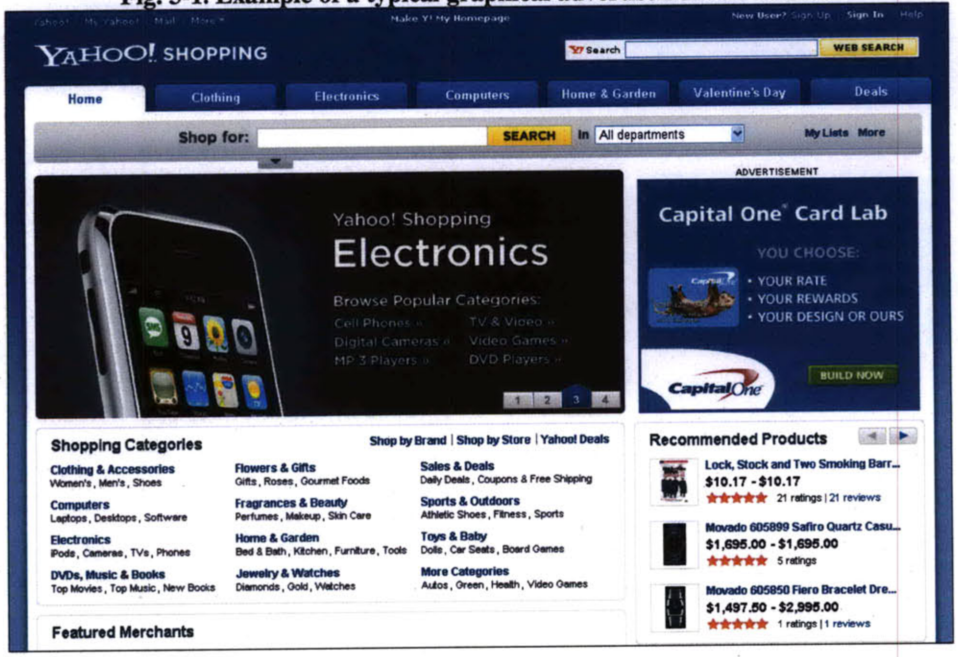


Fig. 3-2. Distribution of customers by age and gender.

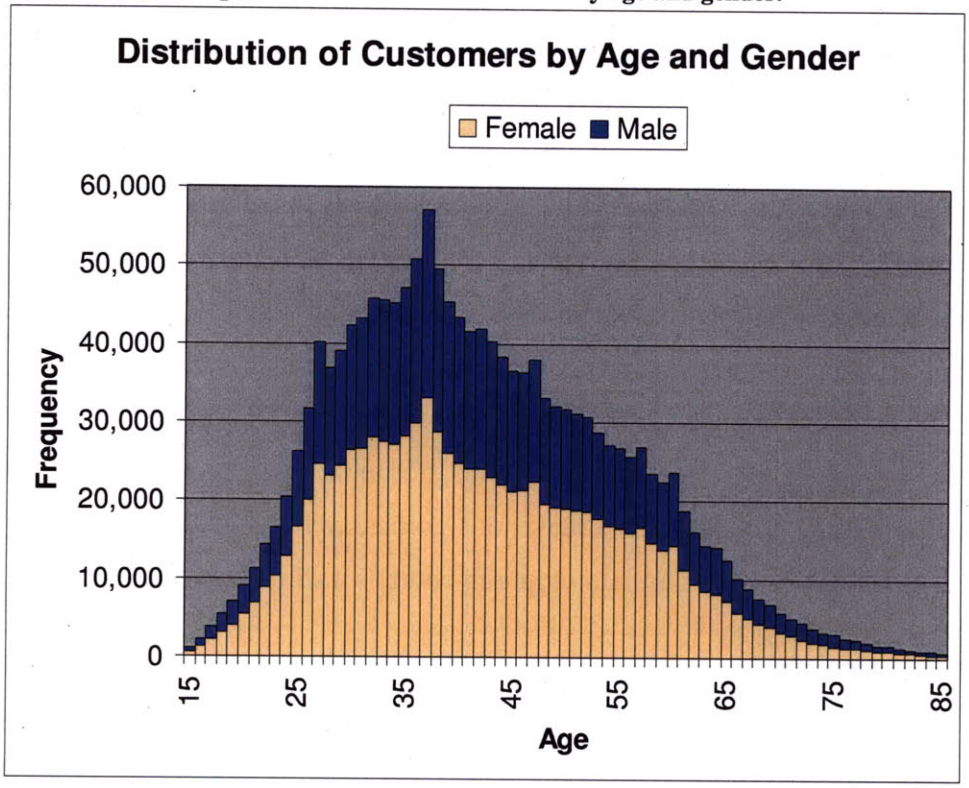


Fig. 3-3. Advertising's effects as a function of customer age.

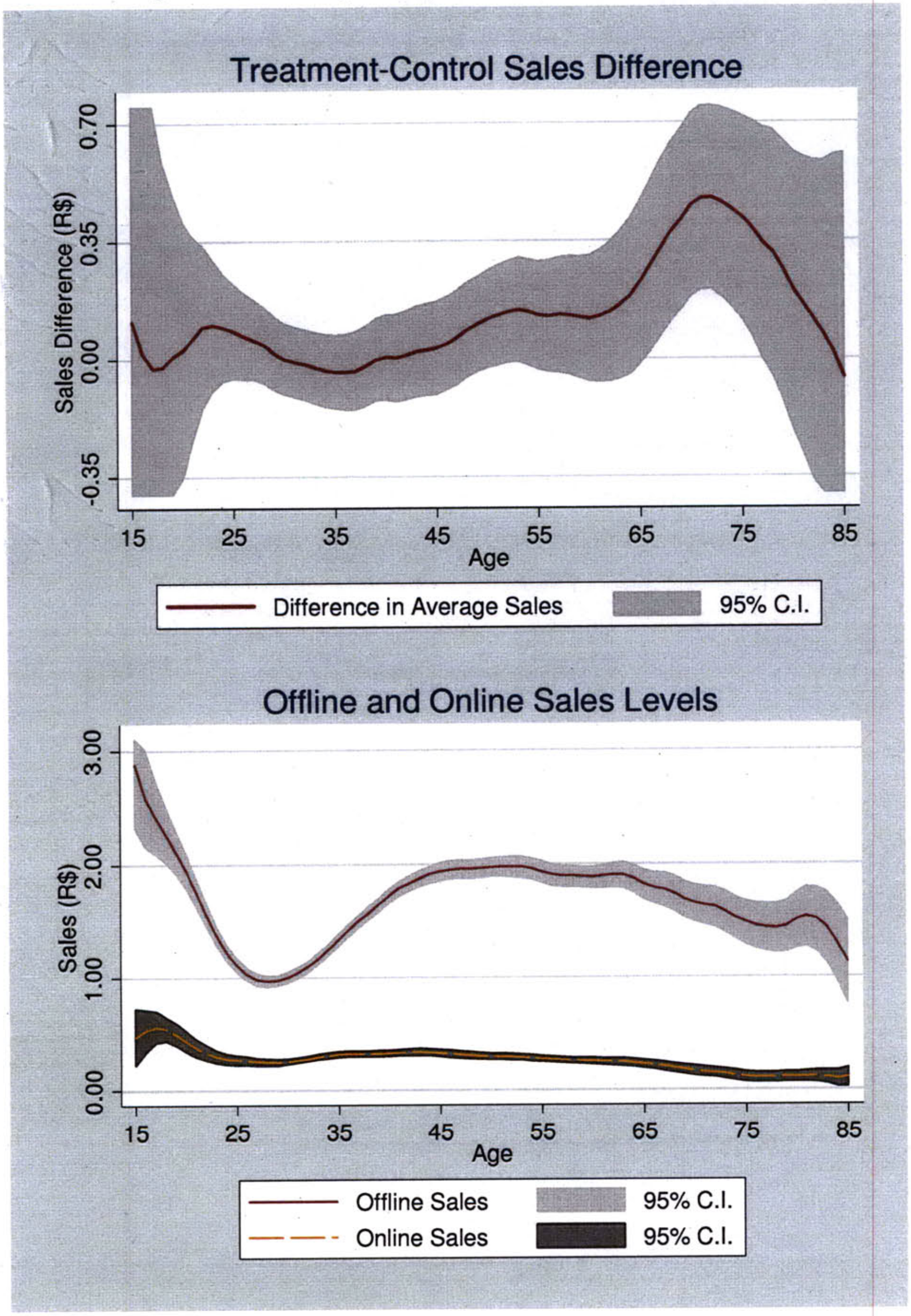


Fig. 3-4. Advertising's effects as a function of customer age, in percentage terms.

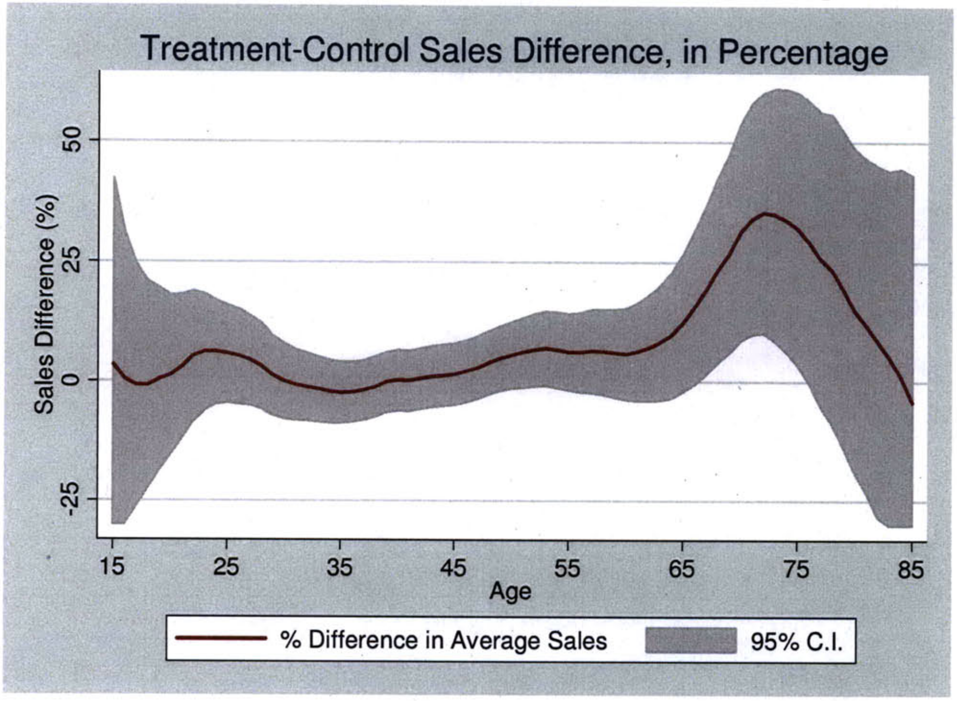
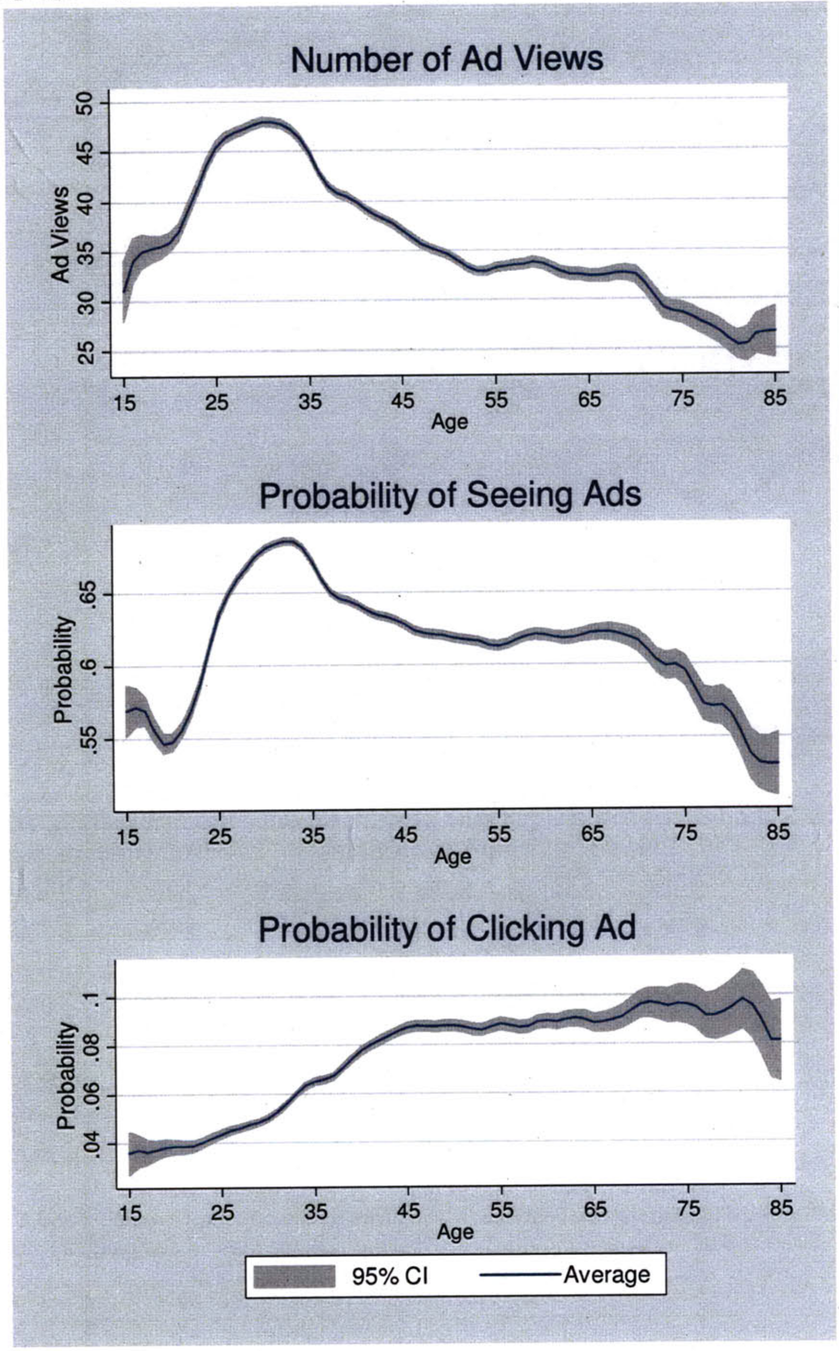


Fig. 3-5. The propensity to view ads declines with age, but the propensity to click increases with age.



IX. Appendix Figures and Tables

Table 3-1. Summary statistics for the three campaigns.

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

Fig. 3-6. Gender and age distribution randomization check.

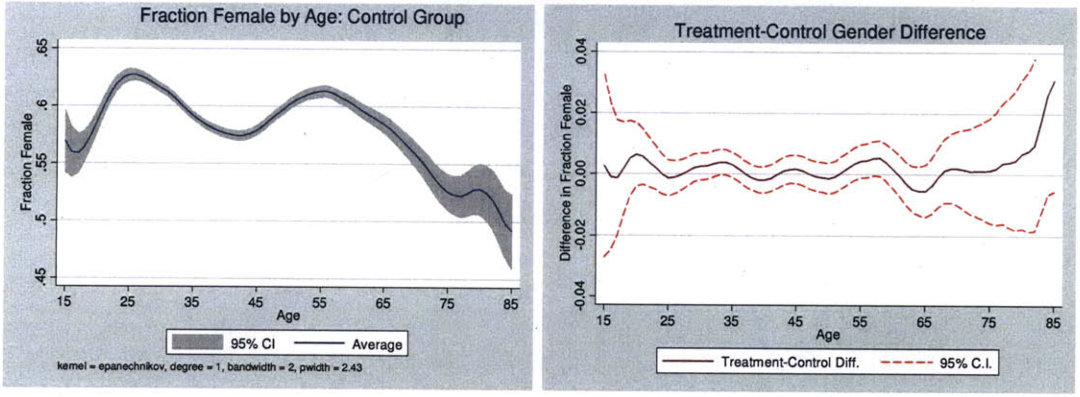


Table 3-2. Ad effects in levels and sales percentages computed by linear regression.

	Effect of Online Ads on Sales*				Ad Effect as % of Sales**		
	Total	Offline	Online	% Offline	Total	Offline	Online
Campaign #1	0.061	0.048	0.014	78%	3.3%	3.1%	4.7%
	(0.037)	(0.035)	(0.013)				
Campaign #2	0.061	0.052	0.009	85%	3.0%	3.0%	2.8%
	(0.044)	(0.042)	(0.013)				
Campaign #3	0.029	0.023	0.006	80%	3.1%	2.9%	3.9%
	(0.028)	(0.026)	(0.008)				
All 3 Campaigns	0.152	0.123	0.029	81%	3.2%	3.0%	3.8%
	(0.069)	(0.064)	(0.022)				

* Each estimate was computed using a regression with sales as the dependent variable and a treatment grou indicator and online and offline sales from the three weeks preceding the first campaign as independent

** Sales levels correspond to the average purchase amount for the control group.

Table 3-3. Ad effects in levels and sales percentages computed by simple treatment-control differences.

	Effect of Online Ads on Sales*				Ad Effect as % of Sales**		
	Total	Offline	Online	% Offline	Total	Offline	Online
Campaign #1	0.053 (0.038)	0.046 (0.035)	0.007 (0.013)	87%	2.9%	3.0%	2.3%
Campaign #2	0.054 (0.044)	0.050 (0.042)	0.004 (0.013)	93%	2.7%	2.9%	1.2%
Campaign #3	0.028 (0.028)	0.023 (0.026)	0.005 (0.008)	82%	2.9%	2.9%	3.3%
All 3 Campaigns	0.134 (0.070)	0.119 (0.064)	0.016 (0.023)	88%	2.8%	2.9%	2.1%

* Each estimate is the difference between the treatment and control group average sales for each category of sales for each campaign.

** Sales levels correspond to the average purchase amount for the control group.

Fig. 3-7. Nonparametric plots of sales versus age for the three weeks preceding the experiment. The average purchasing behavior for each age group validates the randomization. The graphs are oriented in a grid with the three columns representing age group purchases for females, males, and both males and females, respectively, and the three rows representing online, offline, and combined sales. The dark lines are local differences in averages computed by locally linear regression using an Epanechnikov kernel with bandwidth of four years. The dashed lines above and below the difference in averages correspond to asymptotic 95% pointwise confidence intervals.

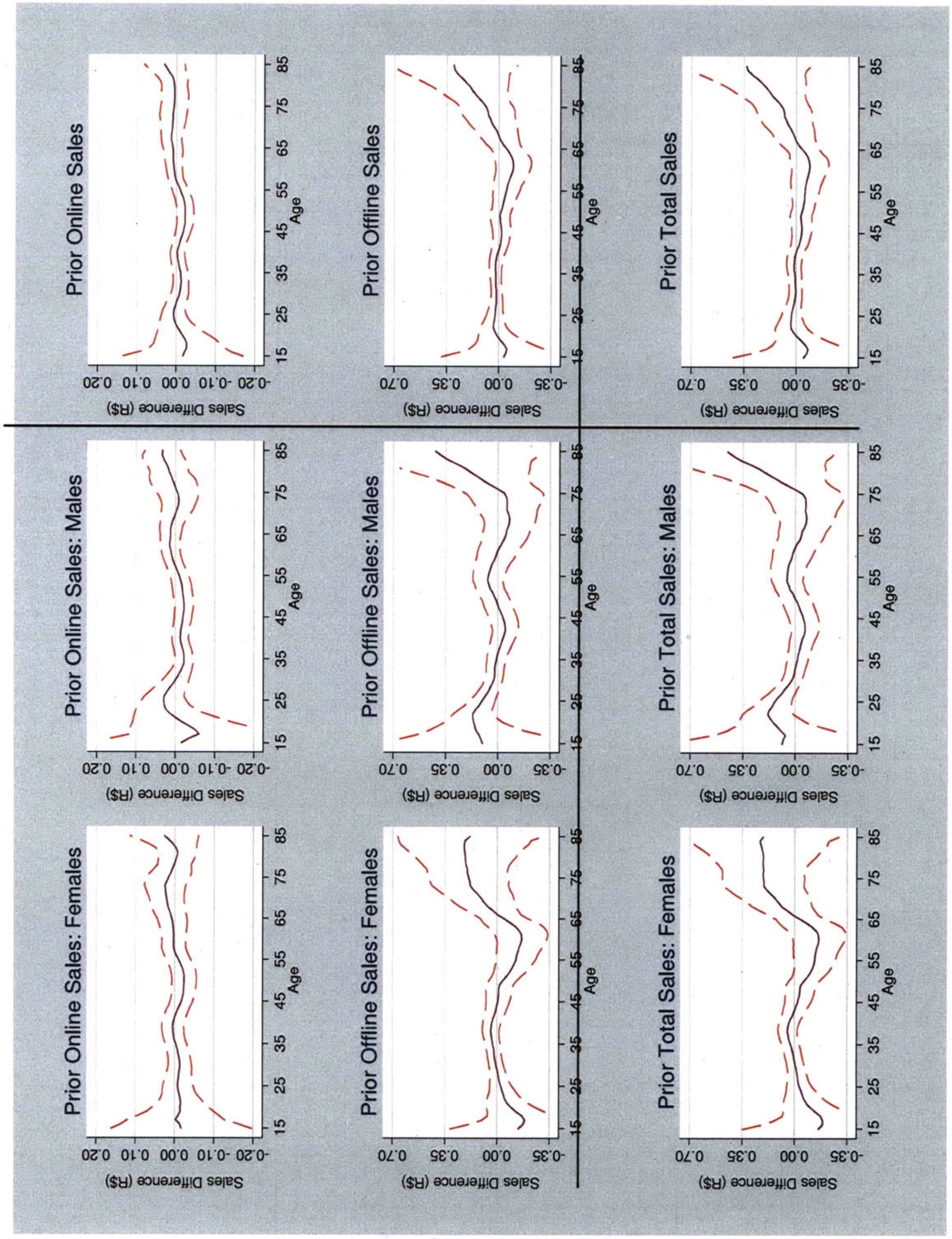


Fig. 3-8. Nonparametric plots of Sales versus age for the two weeks during campaign 1. The average purchasing behavior for each age group shows an effect most pronounced among older women. The graphs are oriented in a grid with the three columns representing age group purchases for females, males, and both males and females, respectively, and the three rows representing online, offline, and combined sales. The dark lines are local differences in averages computed by locally linear regression using an Epanechnikov kernel with bandwidth of four years. The dashed lines above and below the difference in averages correspond to asymptotic 95% pointwise confidence intervals.

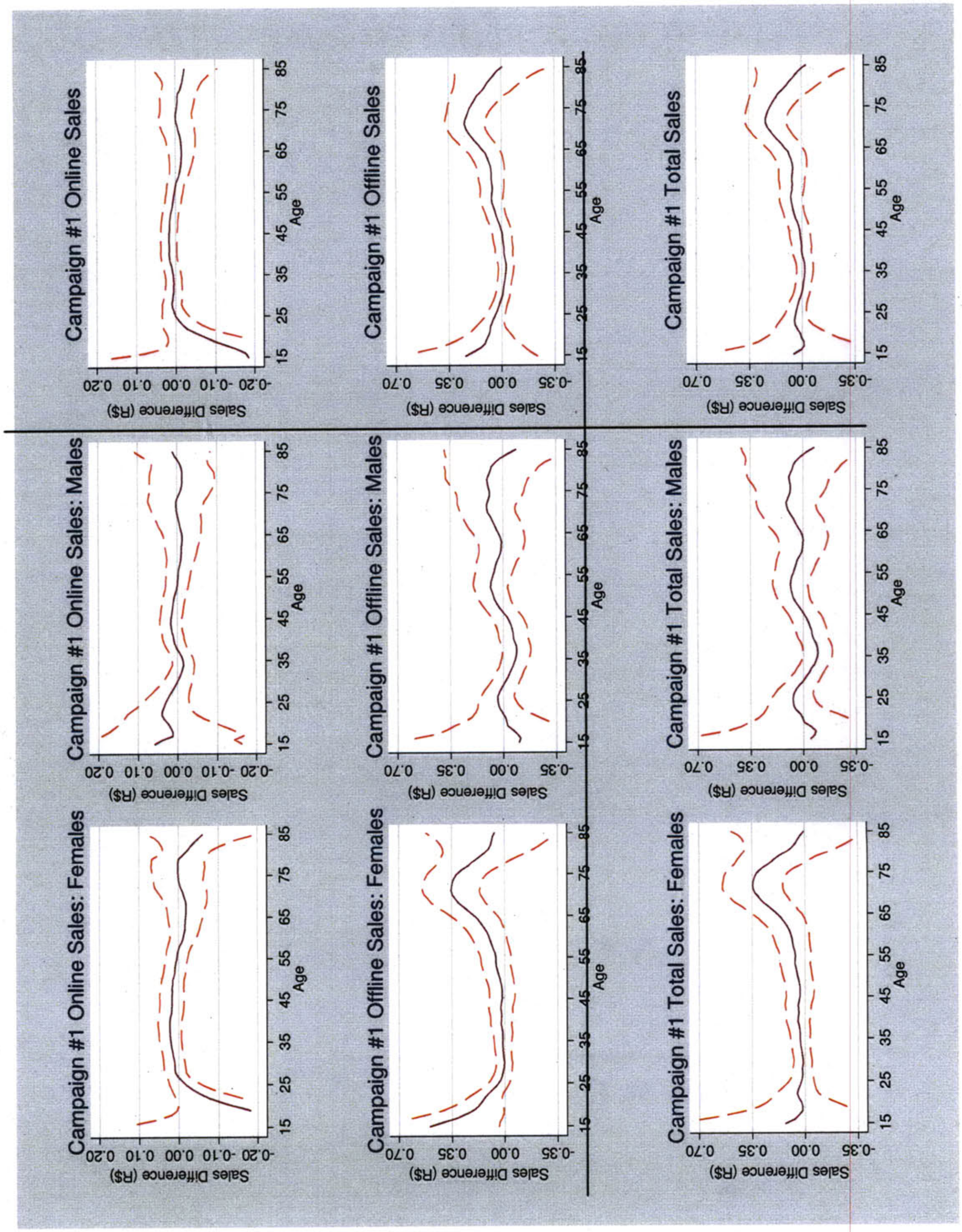


Fig. 3-9. Nonparametric plots of sales versus age for the ten days during campaign 2. The average purchasing behavior for each age group shows a weak effect that is most pronounced among older women. The graphs are oriented in a grid with the three columns representing age group purchases for females, males, and both males and females, respectively, and the three rows representing online, offline, and combined sales. The dark lines are local differences in averages computed by locally linear regression using an Epanechnikov kernel with bandwidth of four years. The dashed lines above and below the difference in averages correspond to asymptotic 95% pointwise confidence intervals.

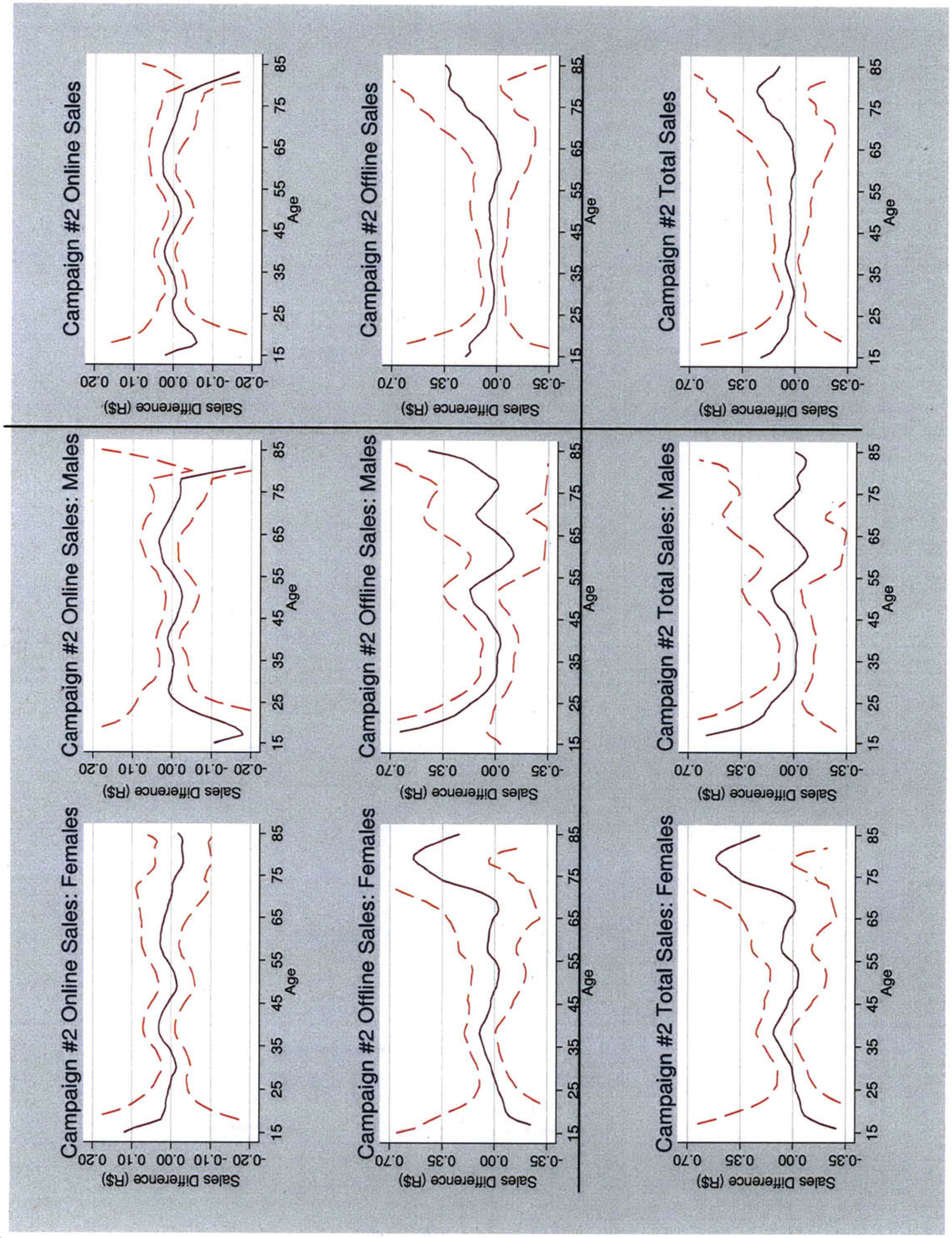


Fig. 3-10. Nonparametric plots of sales versus age for the ten days during campaign 3. The average purchasing behavior for each age group shows a weak effect that is most pronounced among older women. The graphs are oriented in a grid with the three columns representing age group purchases for females, males, and both males and females, respectively, and the three rows representing online, offline, and combined sales. The dark lines are local differences in averages computed by locally linear regression using an Epanechnikov kernel with bandwidth of four years. The dashed lines above and below the difference in averages correspond to asymptotic 95% pointwise confidence intervals.

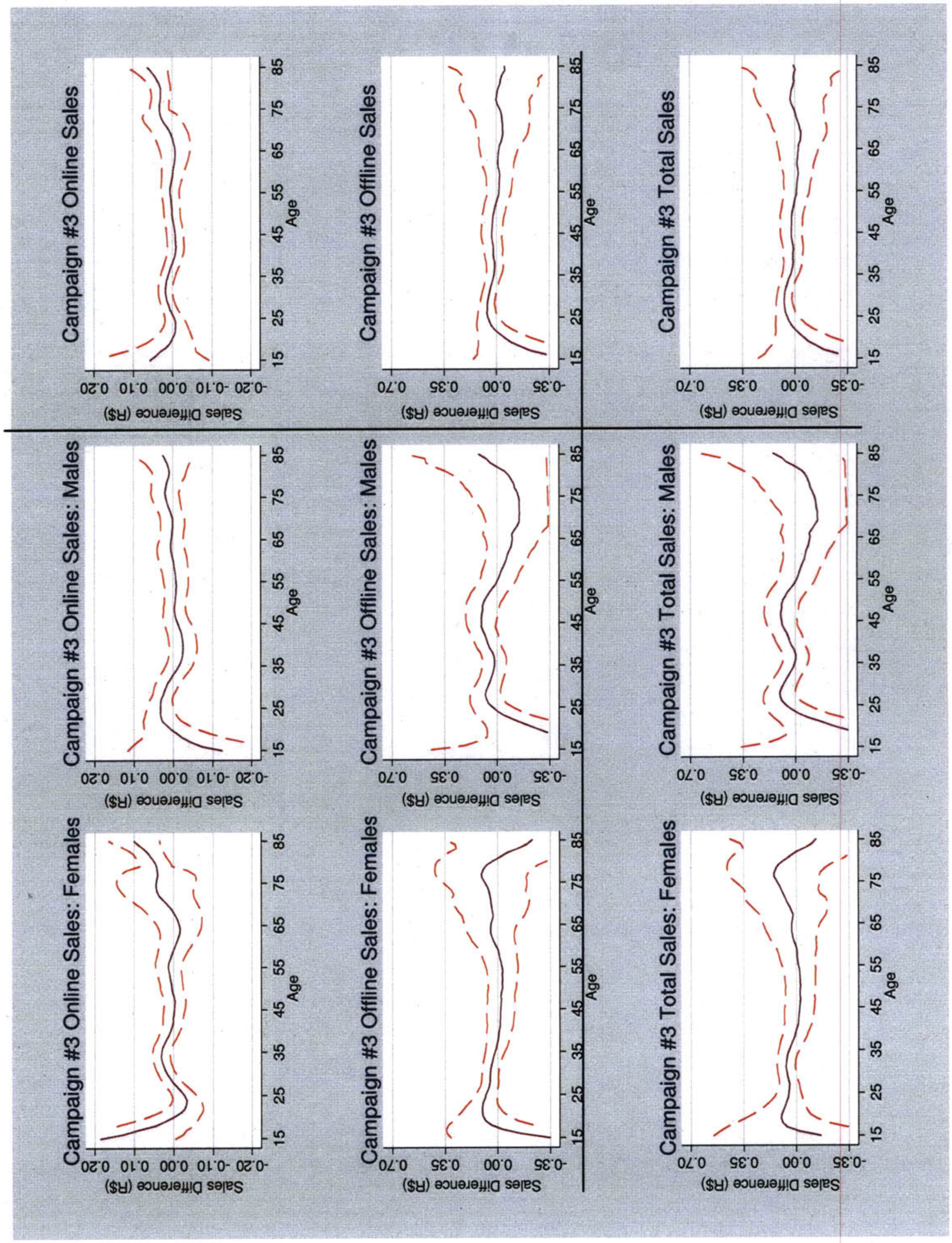
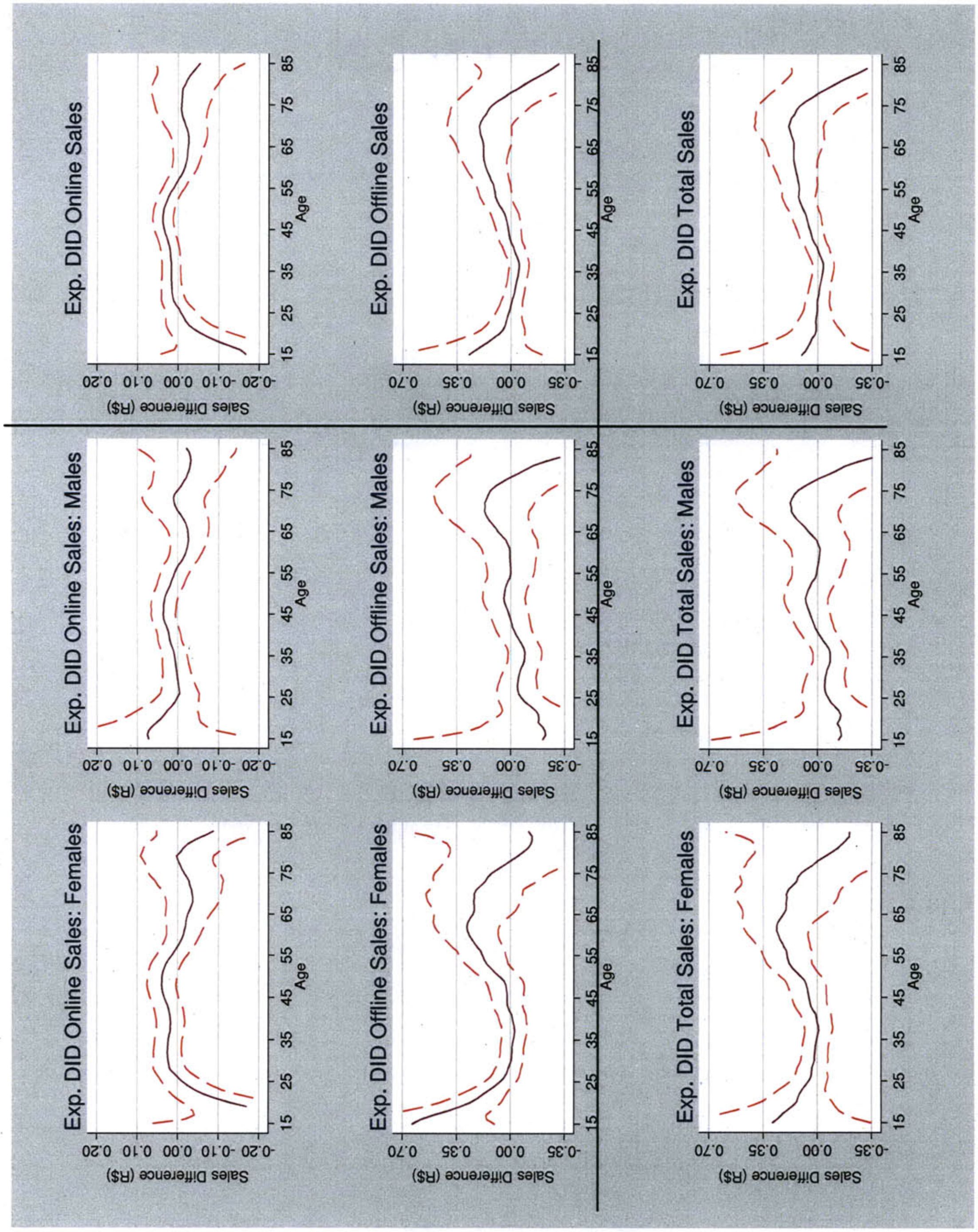


Fig. 3-11. Nonparametric plots of difference between pre-campaign and campaign 1 weekly sales versus age. The average purchasing behavior for each age group shows a strong effect that is most pronounced among older women. The graphs are oriented in a grid with the three columns representing age group purchases for females, males, and both males and females, respectively, and the three rows representing online, offline, and combined sales. The dark lines are local differences in averages computed by locally linear regression using an Epanechnikov kernel with bandwidth of four years. The dashed lines above and below the difference in averages correspond to asymptotic 95% pointwise confidence intervals.



Chapter 4

Does Retail Advertising Work?

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

Randall A. Lewis and David H. Reiley*

Abstract

The effects of online advertising on sales are measured using a randomized experiment performed in cooperation between Yahoo! and a major retailer. Over one million customers are matched between the databases of Yahoo! and the retailer and assigned to treatment and control groups for an online advertising campaign for this retailer. Individual-level data on ad exposure and weekly purchases at this retailer, both online and in stores, are combined and used to find statistically and economically significant impacts of the advertising on sales. The treatment effect persists for weeks after the end of an advertising campaign, and the total effect on revenues is estimated to be more than seven times the retailer's expenditure on advertising during the study. Additional results explore differences in the number of advertising impressions delivered to each individual, online and offline sales, and the effects of advertising on those who click the ads versus those who merely view them.

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

I. Introduction

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

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

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

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

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

In general, advertisers do not systematically vary their levels of advertising to measure the effects on sales.[1] Advertisers often change their levels of advertising over time, as they run discrete "campaigns" during different calendar periods, but this variation does not produce clean data for measuring the effects of advertising because other variables also change concurrently over time. For example, if a retailer advertises more during December than in other months, we do not know how much of the increased sales to attribute to the advertising, and how much to increased holiday demand.

As is well known in the natural sciences, experiments are a great way to establish and measure causal relationships. Randomizing a policy across treatment and control groups allows us to vary advertising in a way that is uncorrelated with all other factors affecting sales, thus eliminating econometric problems of endogeneity and omitted-variable bias. This recognition has become increasingly important in economics and the social science; see Levitt and List (2008) for a summary. We add to this recent literature with an unusually large-scale field experiment involving over one million subjects.

[1] Notable exceptions include direct-mail advertising, and more recently, search-engine advertising, where advertisers do run frequent experiments (on advertising copy, targeting techniques, etc.) in order to measure direct-response effects by consumers. In this study, we address brand advertising, where the expected effects have to do with longer-term consumer goodwill rather than direct responses. In this field, advertising's effects are much less well understood.

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

More recently, Anderson and Simester (2008) experimented with a catalog retailer's frequency of catalog mailings, a direct-mail form of retail advertising. A sample of 20,000 customers received either twelve or seventeen catalog mailings over an eight-month period. When customers received more mailings, they exhibited increased short-run purchases. However, they also found evidence of intertemporal substitution, with the firm's best customers making up for short-run increases in purchases with longer-run decreases in purchases.

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

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

The remainder of this paper is organized as follows. We present the design of the experiment in Section II, followed by a description of the data in Section III. In Section IV, we measure the effect on sales during the first of two[2] advertising campaigns in this experiment. In Section V, we demonstrate and measure the persistence of this effect after the campaign has ended. In Section VI, we examine how the treatment effect of online advertising varies across a number of dimensions. This includes the effect on online versus offline sales, the effect on those who click ads versus those who merely view them, the effect for users who see a low versus high frequency of ads, and the effect on number of customers purchasing versus the size of the average purchase. The final section concludes.

II. Experimental Design

This experiment randomized individual-level exposure to a nationwide retailer's display-advertising campaign on Yahoo! This enabled us to measure the causal effects of the advertising on individuals' weekly purchases, both online and in stores. To achieve this end, we matched the retailer's customer database against Yahoo!'s user database. This match yielded a sample of 1,577,256 individuals who matched on name and either email or postal address.[3]

Of these matched users, we assigned 81% to a treatment group who subsequently viewed two advertising campaigns on Yahoo! from the retailer. The remaining 19% were assigned to the

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

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

control group and saw none of the retailer's ads on Yahoo! The simple randomization was designed to make the treatment-control assignment independent of all other relevant variables.

The treatment group of 1.3 million Yahoo! users was exposed to two different advertising campaigns over the course of two months in fall 2007, separated by approximately one month. Table 4-1 gives summary statistics for the campaigns, which delivered 32 million and 10 million impressions, respectively. By the end of the second campaign, a total of 868,000 users had been exposed to ads. These individuals viewed an average of 48 ad impressions per person.

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

The campaigns in this experiment consisted of "run-of-network" ads on Yahoo! This means that ads appeared on various Yahoo! properties, such as mail.yahoo.com, groups.yahoo.com, and maps.yahoo.com. Figure 4-1 shows a typical display advertisement placed on Yahoo! The large rectangular ad for Netflix[4] is similar in size and shape to the advertisements in this experiment.

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

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

III.Sales and Advertising Data

Table 4-2 provides some summary statistics for the first campaign, providing evidence consistent with a valid randomization.[5] The treatment group was 59.7% female while the control group was 59.5% female, a statistically insignificant difference (p=0.212). The proportion of individuals who did any browsing on the Yahoo! network during the campaign was 76.4% in each group (p=0.537). Even though 76.4% of the treatment group visited Yahoo! during the campaign, only 63.7% of the treatment group actually received pages containing the retailer's ads. On average, a visitor received the ads on only 7.0% of the pages she visited. The probability of being shown an ad on a particular page depends on a number of variables, including user demographics, the user's past browsing history, and the topic of the page visited.

The number of ads viewed by each Yahoo! user in this campaign is quite skewed. The very large numbers in the upper tail are likely due to the activity of non-human "bots," or automated browsing programs. Restricting attention to users in the retail database match should tend to reduce the number of bots in the sample, since each user in our sample has previously made a purchase at the retailer. Nevertheless, we still see a small number of likely bots, with extreme browsing behavior. Figure 4-2 shows a frequency histogram of the number of the retailer's ads viewed by treatment group members that saw at least one of the ads during campaign #1. The majority of users saw fewer than 100 ads, with a mere 1.0% viewing more than 500 ads during the two weeks of the online ad campaign. The maximum number of the ads delivered to a single individual during the campaign was 6050.[6]

One standard statistic in online advertising is the click-through rate, or fraction of ads that were clicked by a user. The click-through rate for this campaign was 0.28%. With detailed user data, we can also tell that the proportion of the designated treatment group who clicked at least

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

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

one ad in this campaign was 4.6% (sometimes called the "clicker rate"). Conditional on receiving at least one ad, the clicker rate was 7.2%.

In order to protect the privacy of individual users, a third party matched the retailer's sales data to the Yahoo! browsing data and anonymized all observations so that neither party could identify individual users in the matched dataset. This weekly sales data includes both online and offline sales and spans approximately 18 weeks: 3 weeks preceding, 2 weeks during, and 1 week following each of the two campaigns. Sales amounts include all purchases that the retailer could link to each individual customer in the database.[7]

Table 4-3 provides a weekly summary of the sales data, while Figure 4-3 decomposes the sales data into online and offline components. We see that offline (in-store) sales represent 86% of the total. Combined weekly sales are quite volatile, even though averaged across 1.6 million individuals, ranging from less than R$0.60 to more than R$1.60 per person. The standard deviation of sales across individuals is much larger than the mean, at approximately R$14. The mean includes a large mass of zeroes, as fewer than 5% of individuals in a given week make any transaction (see last column of Table 4-3). For those who do make a purchase, the transaction amounts exhibit large positive and negative amounts (the latter representing returns), but well over 90% of purchase amounts lie between –R$100 and +R$200.

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

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

IV. Basic Treatment Effect in Campaign #1

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

Table 4-4 gives initial results comparing sales between treatment and control groups. We look at total sales (online and offline) during the two weeks of the campaign, as well as total sales during the two weeks prior to the start of the campaign. During the campaign, we see that the treatment group purchased R$1.89 per person, compared to the control group at $1.84 per person. This suggests a positive average treatment effect (intent to treat) from ad exposures of R$0.053 (0.038) per person, but the effect is not statistically significant at conventional levels (p=0.162).

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

Because only 64% of the treatment group was actually treated with ads, this simple treatment-control comparison has been diluted with the 36% of individuals who did not see any ads during this campaign. (Recall that they did not see ads because of their individual browsing choices.) Ideally, we would remove these 36% of individuals both from the treatment and control groups in order to get an estimate of the advertising treatment on those who could be treated. Unfortunately, we are unable to observe which control-group members would have seen ads for this campaign had they been in the treatment group.[8] Instead of removing these individuals, we

[8] We recorded zero impressions of the retail ad campaign to every member of the control group, which makes it impossible to distinguish those control group members who would have seen ads. The Yahoo! ad server uses a

scale up our diluted treatment effect (R$0.05) by dividing by 0.64, the fraction of individuals treated.[9] This gives us an estimate of the treatment effect on those treated with ads: R$0.083 (0.059). The standard error is also scaled proportionally, leaving the level of statistical significance unaffected (p=0.162).

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

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

complicated set of rules and constraints to determine which ad will be seen by a given individual on a given page. For example, a given ad might be shown more often on Yahoo! Mail than on Yahoo! Finance. If another advertiser has targeted females under 30 during the same time period, then this ad campaign may have been relatively more likely to be seen by other demographic groups. Our treatment-control assignment represented an additional constraint. Because of the complexity of the server delivery algorithm, we were unable to model the hypothetical distribution of ads delivered to the control group with an acceptable level of accuracy, so we cannot restrict attention to treated individuals without incurring significant selection bias in our estimate of the treatment effect.

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

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

$$SawAds_{i,t} = \pi_{0,t} + \pi_{1,t} Treatment_i + \eta_{i,t}$$

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

behavior negatively correlated with Yahoo! browsing and ad delivery, and the large bias that can result from cross-sectional study of advertising exposure in the absence of an experiment.

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

During the campaign, there persists a sales difference between treated and untreated members of the treatment group, but this difference becomes smaller. While untreated individuals' sales drop by R$0.10 from before the campaign, treated individuals' sales remained constant. (Control-group mean sales also fell by R$0.10 during the same period, just like the untreated portion of the treatment group.) This suggests that advertisements may be preventing treated individuals' sales from falling like untreated individuals' sales did. This will lead us to our preferred estimator below, a difference in differences between treated and untreated individuals (where "untreated" pools together both control-group members and untreated members of the designated treatment group).

Before presenting our preferred estimator, we first look at the shape of the distribution of sales. Figure 4-4 compares histograms of sales amounts for the treatment group and control group, omitting those individuals for whom there was no transaction. For readability, these histograms exclude the most extreme outliers, trimming approximately 0.5% of the positive purchases from both the left and the right of the graph.[10] Relative to the control, the treatment density has less mass in the negative part of the distribution (net returns) and more mass in the positive part of the distribution. These small but noticeable differences both point in the direction of a positive treatment effect, especially when we recall that this diagram is diluted by the 34% of customers who did not browse enough to see any ads on Yahoo! Figure 4-5 plots the

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

difference between the two histograms in Figure 4-4. The treatment effect is the average over this difference between treatment and control sales distributions.

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

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

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

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

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

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

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

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

Using difference in differences, the estimated average treatment effect of being treated by viewing at least one of the retailer's ads during the campaign is R$0.102 with a standard error of R$0.043. This effect is statistically significant ($p<0.01$) as well as economically significant, representing an average increase of 5% on treated individuals' sales. Based on the 814,052 treated individuals, the estimate implies an increase in revenues for the retailer of R$83,000 ± 68,000 (95% confidence interval) due to the campaign. Because the cost of campaign #1 was approximately R$25,000,[11] the point estimate suggests that the ads produced more than 325% as much revenue as they cost the retailer. We conclude that retail advertising does, in fact, work.

The main identifying assumption of the DID model is that each individual's idiosyncratic tendency to purchase from the retailer is constant across time, and thus the treatment variable is uncorrelated with the DID error term. This assumption could be violated if some external event at some point during the experiment had different effects on the retail purchases of those who did and did not see ads. For example, perhaps in the middle of the time period studied, the retailer did a direct-mail campaign we do not know about, and the direct mail was more likely to reach those individuals in our study who browsed less often on Yahoo! Fortunately, our previous experimentally-founded estimates are very similar in magnitude to the DID estimates: R$0.083 for the simple comparison of levels between treatment and control, versus R$0.102 for the DID estimate.

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

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

V. Persistence of the Effects

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

A. Sales in the Week after the Campaign Ended

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

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

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

B. More than One Week after the Campaign Ended

Could the effects be persistent even more than a week after the campaign ends? We investigate this question using sales data collected for purposes of evaluating campaign #2. Recall that for both campaigns, we obtained three weeks of sales data before the start of the campaign, and three weeks of sales data after the start of the campaign. It turns out that the earliest week of pre-campaign sales for campaign #2 happens to be the fourth week after the start of campaign #1, so we can use that data to examine the treatment effect of campaign #1 in its fourth week.[12]

In order to check for extended persistence of advertising, we use the same DID model as before, estimated on weekly sales. Our "pre-period" sales will be the weekly average of sales in the three weeks preceding the start of campaign #1. Our "post-period" sales will be the sales during a given week after the start of campaign #1. We then compute a separate DID estimate for each week, beginning with the first week of campaign #1 and ending with the week following campaign #2.[13]

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

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

Table 4-5 displays the results, and Figure 4-8 represents them graphically. In the figure, vertical lines indicate the beginning (solid) and end (dashed) of both campaigns. 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$0.036, with individual weekly estimates ranging from R$0.004 to R$0.061. Some of the individual weekly treatment effects are statistically indistinguishable from zero (95% confidence intervals graphed in Figure 4-8), but, strikingly, every single point estimate is positive.[14] We particularly note the large, positive effects estimated during the inter-campaign period, more than three weeks after ads stopped showing for the retailer's first campaign on Yahoo!

To obtain an estimate of the cumulative effect of both campaigns, we use all nine weeks of data. We present the results of two different methods in Table 4-6. The first method estimates the average weekly treatment effect by comparing the average of the nine weeks after campaign #1, scaled according to the number of weeks accounted for (eight weeks, three days) to the average of the three weeks prior to campaign #1. The estimate is R$0.037. We then multiply this estimate by the number of independent weeks of sales data in the sample period, which is actually eight weeks.[15] This estimate gives us an average treatment effect of the ads (on those who saw at least one ad during one of the campaigns) of R$0.299 with a standard error of R$0.123.

A more econometrically efficient method is to compute an average of the nine weekly estimates of the treatment effect, taking care to report standard errors that account for the covariances between regression coefficients across weeks. Table 4-6 reports an optimally weighted average of the nine per-week treatment effects,[16] with a simple average included for

[14] We find this fact striking, but in order to avoid overstating its significance, we note that the weekly estimates are not independent of each other. Each week's DID estimator relies on the same three weeks of pre-campaign data.

[15] Note that the first of three weeks prior to the start of campaign #2 overlaps with the week following campaign #1 for three days (see footnote 12) and that one of campaign #2's second "week" is actually only three days, since the campaign was only ten days long (see footnote 13).

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

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

where the betas are from least squares regression coefficients from regressing Y_j on X_j and Y_k on X_k. We estimate $Cov(\varepsilon_1,\varepsilon_2)$ from the residuals of each regression. One could use the simple estimator

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

comparison. The weighted average is R$0.035 (R$0.0155). We then multiply this number by eight to get a total effect across the entire time period of observation, since the "nine-week" time period actually includes a total of only eight weeks. This multiplication gives us R$0.311 (R$0.117). This is slightly larger than the total effect reported in the previous paragraph, because it assumes that all users were treated from the beginning of campaign #1.

To estimate the total benefit of the two campaigns, we take our estimate of R$0.311 and multiply it by the average number of users who had already been treated with ads in a given week, which turns out to be 816,000. This gives us a 95% confidence interval estimate of the total incremental revenues due to ads of R$253,000 ± 188,000. For comparison, the total cost of these advertisements to the advertiser was R$33,000. Thus, our point estimate says that the total revenue benefit of the ads was nearly eight times the cost of the campaign, while the lower bound of our 95% confidence interval indicates a benefit of two times the cost. Even more strongly than before, we conclude that retail advertising works!

Similar to the specification test computed for the DID estimate during the first 3 weeks, we perform a specification test for each of the weekly DID estimates. This test determines whether the control group and the untreated members of the treatment group might pursue different time-varying purchasing behavior, which would invalidate our DID estimator's strategy of pooling these two groups. We present the results of the weekly estimates of this difference in Figure 4-9. During each of the 9 weeks following the start of campaign #1, the difference in time-series differences between control and untreated treatment group members fails to reject the null hypothesis that the DID model is correctly specified. Our failure to find any significant time-varying deviations from zero suggests that two-month extensions of DID can be reliable even when the pre-period is short and distant in time from the treatment period.

C. Summary of Persistence Results

To summarize the main result of this section, we find that the retail image advertising in this experiment led to persistent positive effects on sales for a number of weeks after the ads

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

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

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

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

VI. Detailed Results for Campaign #1

In this section, we dig deeper into several other dimensions of the data for the first campaign. First, we decompose the effects of online advertising into offline versus online sales, showing that more than 90% of the impact is offline. We also demonstrate that most of the substantial impact on in-store sales occurs for users who merely view the ads but never click them. Second, we examine how the treatment effect varies with the number of ads viewed by each user. Third, we decompose the effects of online advertising into the probability of a transaction versus the size of the purchase conditional on a transaction. We perform all of these analyses only on campaign #1. The second campaign cannot be analyzed cleanly on its own because the treatment and control groups were not re-randomized for campaign #2 and the ad effects may be persistent as shown in the previous section. For the results in this section, we use our preferred specification of a difference in differences for three weeks before and after the start of campaign #1, comparing treated versus untreated individuals.

A. Offline versus Online Sales and Views versus Clicks

In Table 4-7, we present a decomposition of the treatment effect into offline and online components by running the previous difference-in-differences analysis separately for offline and online sales. The first line of the table shows that the vast majority of the treatment effect comes from brick-and-mortar sales. The treatment effect of R\$0.166 per treated individual turns out to consist of a R\$0.155 effect on offline sales plus a R\$0.011 effect on online sales. In other words, 93% of the treatment effect of the online ads occurred in *offline* sales. This result will be surprising to those who assume that online ads have an impact only on online sales.

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

We answer this question in the second and third lines in Table 4-7. We partition the set of treated individuals into those who clicked on an ad (line 2) versus those who merely viewed ads but did not click any of them (line 3). Of the 814,000 individuals treated with ads, 7.2% clicked on at least one ad, while 92.8% merely viewed them. With respect to total sales, we see a treatment effect of R$0.139 on those who merely view ads, and a treatment effect of R$0.508 on those who click them. Our original estimate of the treatment effect can be decomposed into the separate effects for viewers versus clickers, using their relative weights in the population: R$0.166 = (92.8%)(R$0.139) + (7.2%)(R$0.508). The first component—the effect on those who merely view but do not click ads—represents 78% of the total treatment effect. Thus clicks, though the standard performance measure in online advertising, fail to capture the vast majority of the effects on sales.

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

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

We saw in Figure 4-2 that different individuals viewed very different numbers of ads during campaign #1. We now ask how the treatment effect varies with the number of ads viewed.

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

Figure 4-10 gives the results, together with 95% confidence-interval bands around the conditional mean. We see that the treatment effect is initially increasing in the number of ads viewed. The effect peaks at approximately 50 ads viewed, for a maximum treatment effect of R$0.25, and remains almost flat at this level until it reaches 100 ad impressions per person. Beyond this point, the data becomes so sparse (only 6.1% of the treatment group sees more than 100 ad views) that the effect is no longer statistically distinguishable from zero.

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

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

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

C. Probability of Purchase versus Basket Size

So far, our analysis has focused on advertising's effects on the average purchase amount per person, including those who made purchases as well as those who did not. We can decompose the effects of advertising into two separate channels of interest to retailers: the effect on the probability of a transaction, and the effect on "basket size," or purchase amount conditional on a transaction. To provide some base numbers for reference, during the three-week period after the start of the campaign, individuals treated with ads had a 6.48% probability of a transaction, and the average basket size was R$40.72 for those who purchased. The product of these two numbers gives the average (unconditional) purchase amount of R$2.64 per person. We reproduce these numbers in the last column of Table 4-8 for comparison to our treatment-effect results.

In Table 4-8, the first column shows the estimates of the treatment effects on each variable of interest. As before, our treatment effects come from a difference in differences, comparing those treated with ads versus those untreated, using three weeks of data before and after the start of the campaign.

First we investigate advertising's impact on the probability of a transaction,[17] with results shown in the first row of the table. We find an increase of 0.102% in the probability of purchase as a result of the advertising, and the effect is statistically significant (p=0.03). This represents an increase of approximately 1.6% relative to the average probability of a transaction.

Next, we consider the effect on basket size. In this application, we wish to analyze the magnitude of a purchase conditional on a transaction. Since sales data are sparse and most purchasers do not purchase in both time periods, we cannot employ the same DID estimator as before, running a regression with the dependent variable of time differences in sales by individual. Instead, we compute DID using group means of basket size, and to compute a

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

consistent standard error we pay careful attention to possible time-series correlation.[18] As shown in the second row of Table 4-8, the advertising campaign produced an increase in basket size of R$1.75, which is statistically significant (p=0.018). Compared with the baseline basket size of $40.72, this represents an increase of approximately 4.5%.

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

VII. Conclusion

Despite the economic importance of the advertising industry, the effects of advertising on sales have been extremely difficult to quantify. In this study, we take a substantial step forward in this measurement problem by conducting a large-scale field experiment that systematically varies advertising to a subject pool of over one million retail customers on Yahoo! The randomized experiment allows us to measure causal effects of advertising on sales. The simple treatment-control differences are rather noisy, because sales at this retailer have high variance and this online advertising campaign is just one of many factors that influence purchases. For more precise estimates, we employ a difference-in-difference estimator using panel data on weekly individual transactions, exploiting both experimental and non-experimental variation in advertising exposure. This DID estimator requires more assumptions than the simple treatment-control difference estimator, but the two estimators provide similar results and the DID estimator passes a specification test on the assumptions.

Our primary result is that retail advertising works! We find positive, sizeable, and persistent effects of online retail advertising on retail sales. The ad effects appear to persist for several weeks after the last ad has been shown. In total, we estimate that the retailer gained

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

incremental revenues more than seven times as large as the amount it spent on the online ads in this experiment.

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

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

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

Another important result is a demonstration of how poorly one can measure the causal effects of advertising using common modeling strategies. If we had neither an experiment nor panel data available to us, but instead attempted to estimate these effects using cross-sectional variation in endogenous advertising exposure, we would have obtained a result that was opposite in sign to the true estimate. The magnitude of the selection bias would be more than three times the magnitude of the true measured effect of advertising.

In future research, we hope to replicate these results with other retailers. We are using what we have learned in this study in order to design better experiments: for example, future experiments will carefully mark control-group members who would not have browsed in ways that exposed them to ads, so that we can more efficiently estimate the treatment effect on the treated in the experiment. We also wish to investigate related factors in online advertising, such as the value of targeting customers with particular demographic or online-browsing-behavior attributes that an advertiser may think desirable. The ability to conduct a randomized experiment with a million customers and to match individual-level sales and advertising data makes possible exciting new measurements about the economic effects of advertising, and we look forward to additional explorations on this new frontier.

VIII. References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

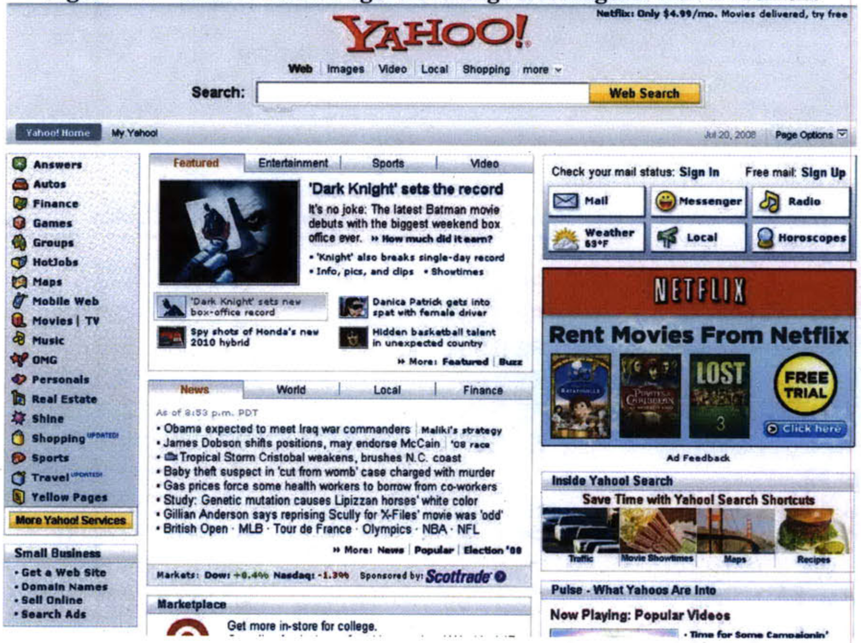
Schmalensee, Richard. 1972. *The Economics of Advertising*. Amsterdam: North-Holland.

IX. Tables and Figures

Table 4-1 - Summary Statistics for the Campaigns

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

Figure 4-1– Yahoo! Front Page with Large Rectangular Advertisement



- Get a Web Site
- Domain Names
- Sell Online
- Search Ads

'Dark Knight' sets the record

It's no joke: The latest Batman movie debuts with the biggest weekend box office ever. » How much did it earn?

- 'Knight' also breaks single-day record
- Info, pics, and clips
- Showtimes

'Dark Knight' sets new box-office record

Danica Patrick gets into spat with female driver

Spy shots of Honda's new 2010 hybrid

Hidden basketball talent in unexpected country

» More: Featured | Buzz

As of 8:53 p.m. PDT

- Obama expected to meet Iraq war commanders | Maliki's strategy
- James Dobson shifts positions, may endorse McCain | '08 race
- Tropical Storm Cristobal weakens, brushes N.C. coast
- Baby theft suspect in 'cut from womb' case charged with murder
- Gas prices force some health workers to borrow from co-workers
- Study: Genetic mutation causes Lipizzan horses' white color
- Gillian Anderson says reprising Scully for 'X-Files' movie was 'odd'
- British Open · MLB · Tour de France · Olympics · NBA · NFL

» More: News | Popular | Election '08

Markets: Dow: +0.4% Nasdaq: -1.3% Sponsored by: Scottrade

Get more in-store for college.

- Time for Some Campaignin'

Table 4-2 - Basic Summary Statistics for Campaign #1

	Control	Treatment
% Female	59.5%	59.7%
% Retailer Ad Views > 0	0.0%	63.7%
% Yahoo Page Views > 0	76.4%	76.4%
Mean Y! Page Views per Person	358	363
Mean Ad Views per Person	0	25
Mean Ad Clicks per Person	0	0.056
% Ad Impressions Clicked (CTR)	-	0.28%
% People Clicking at Least Once	-	4.59%

Figure 4-2 - Ad Views Histogram

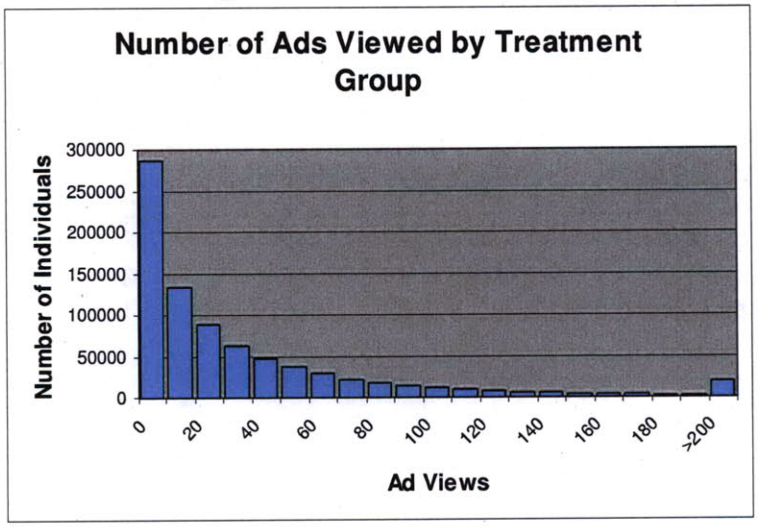


Table 4-3 - Weekly Sales Summary

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

N=1,577,256 observations per week

Figure 4-3 - Offline and Online Weekly Sales

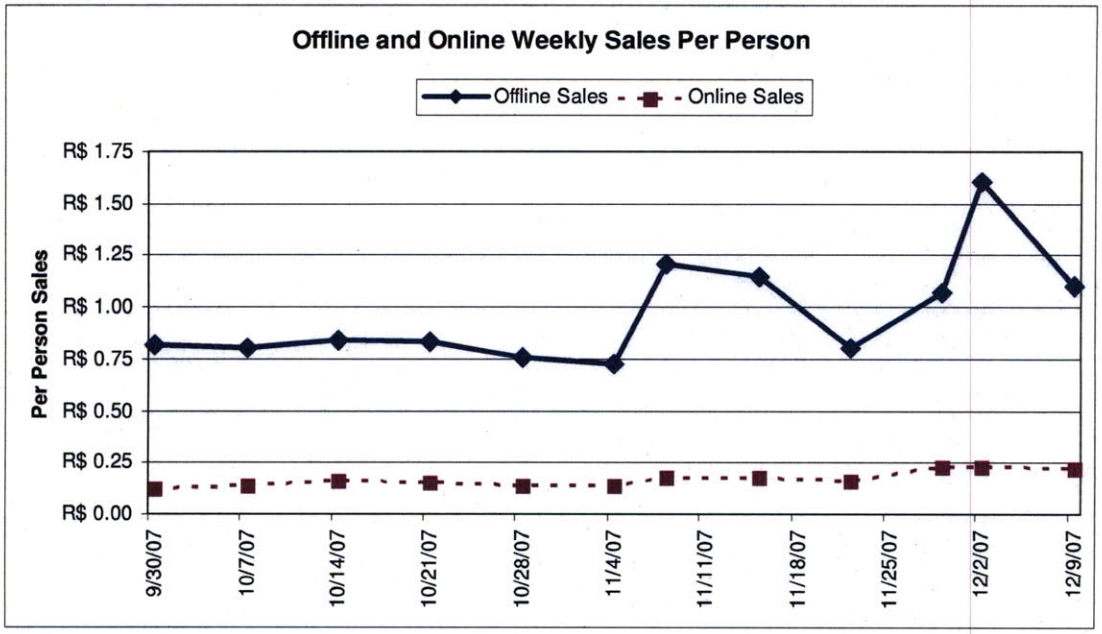


Table 4-4 - Two Week Treatment Effect Offline/Online Decomposition

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

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

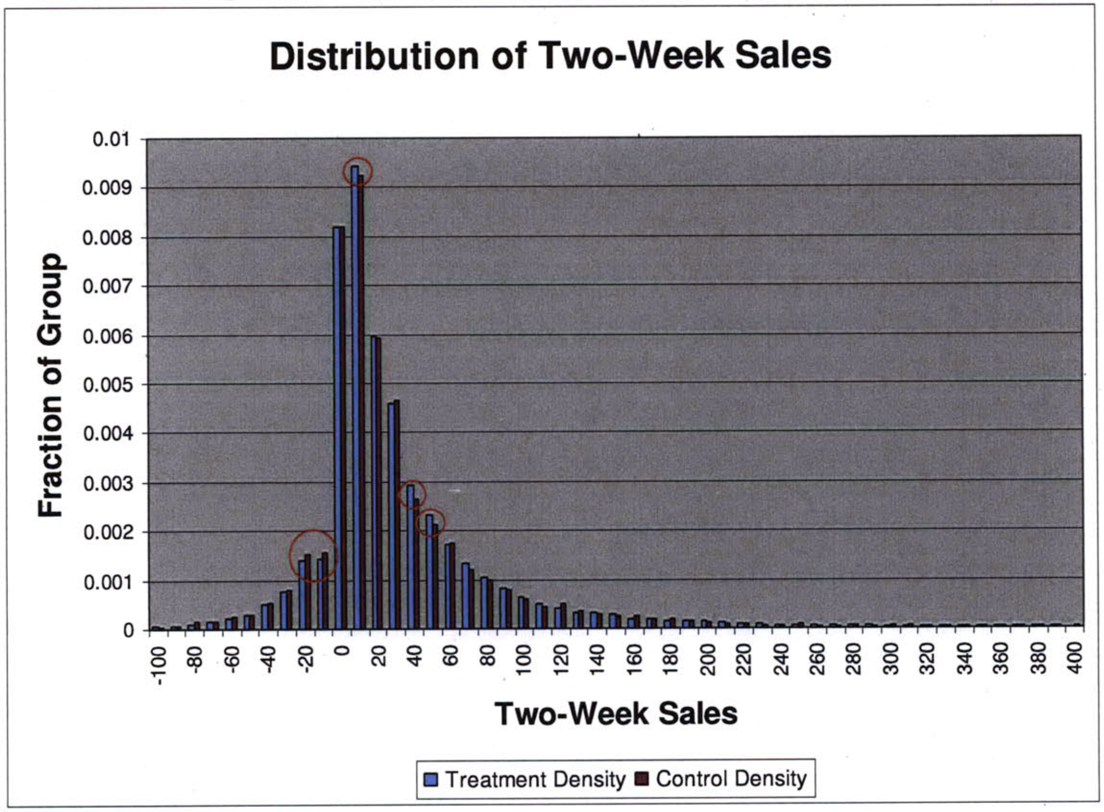


Figure 4-5 - Difference between Treatment and Control Sales Histograms

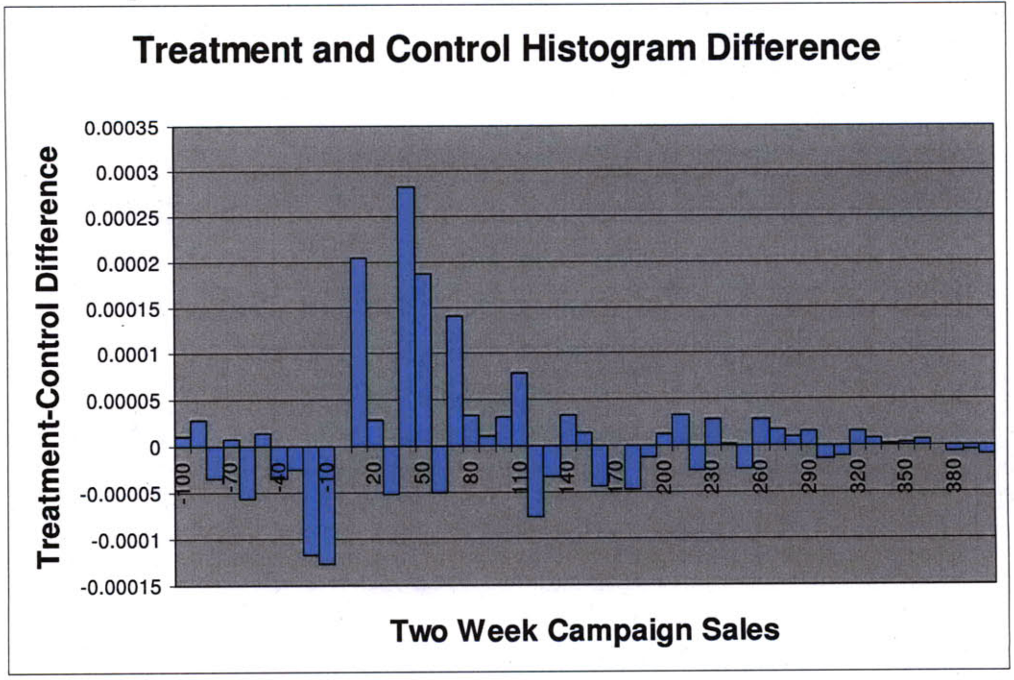


Figure 4-6 - Histogram of Difference in Three-Week Sales for Treated and Untreated Groups

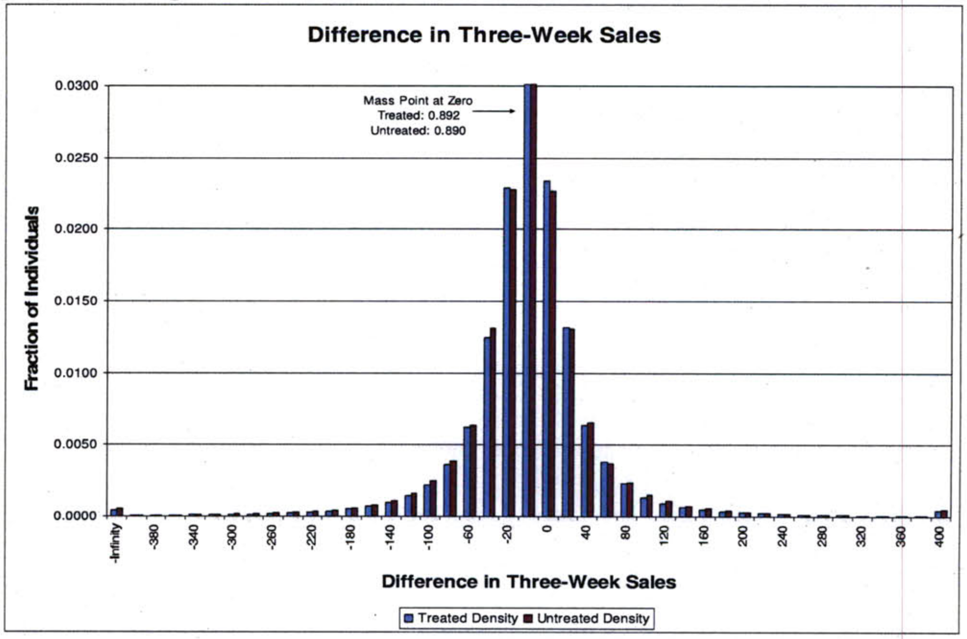
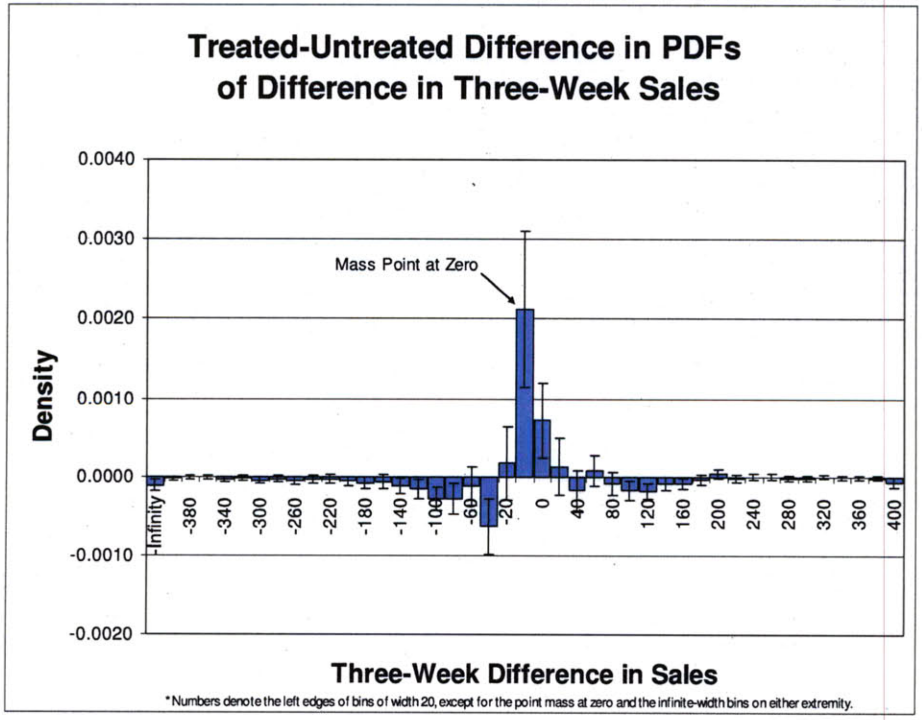


Figure 4-7 - Difference in Treated and Untreated Three-Week Sales Histograms



*Numbers denote the left edges of bins of width 20, except for the point mass at zero and the infinite-width bins on either extremity.

Table 4-5 - Weekly Summary of Effect on the Treated

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

N=1,577,256 obs. per week

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

Figure 4-8 - Weekly DID Estimates of the Treatment Effect for Both Campaigns

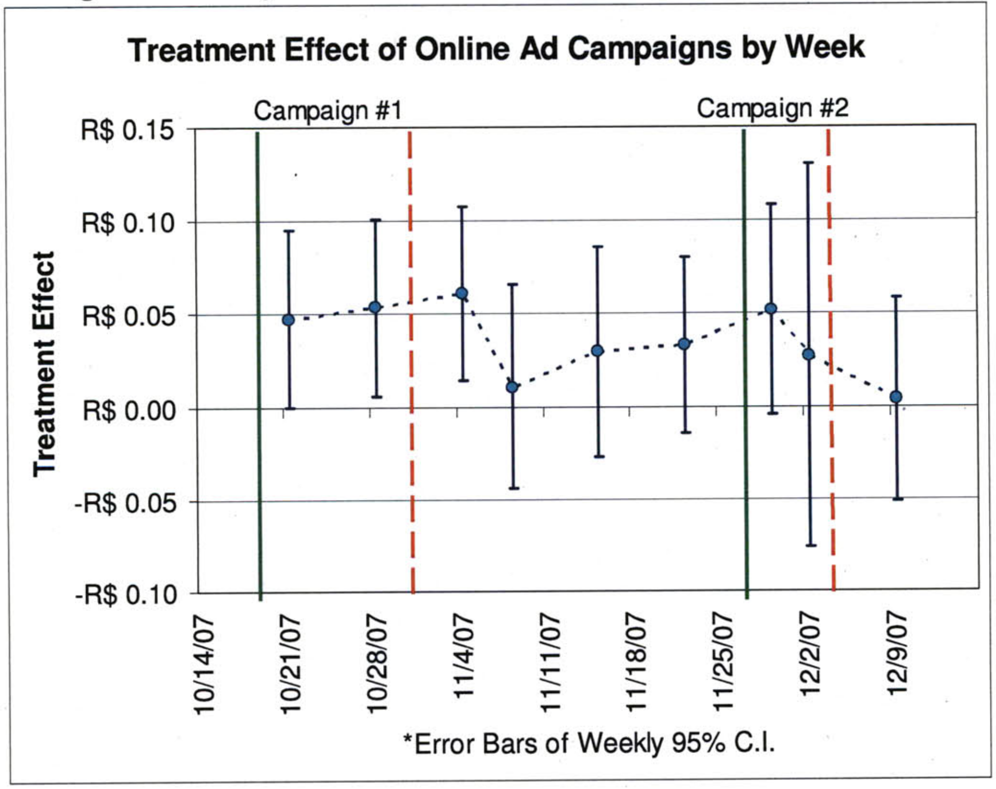
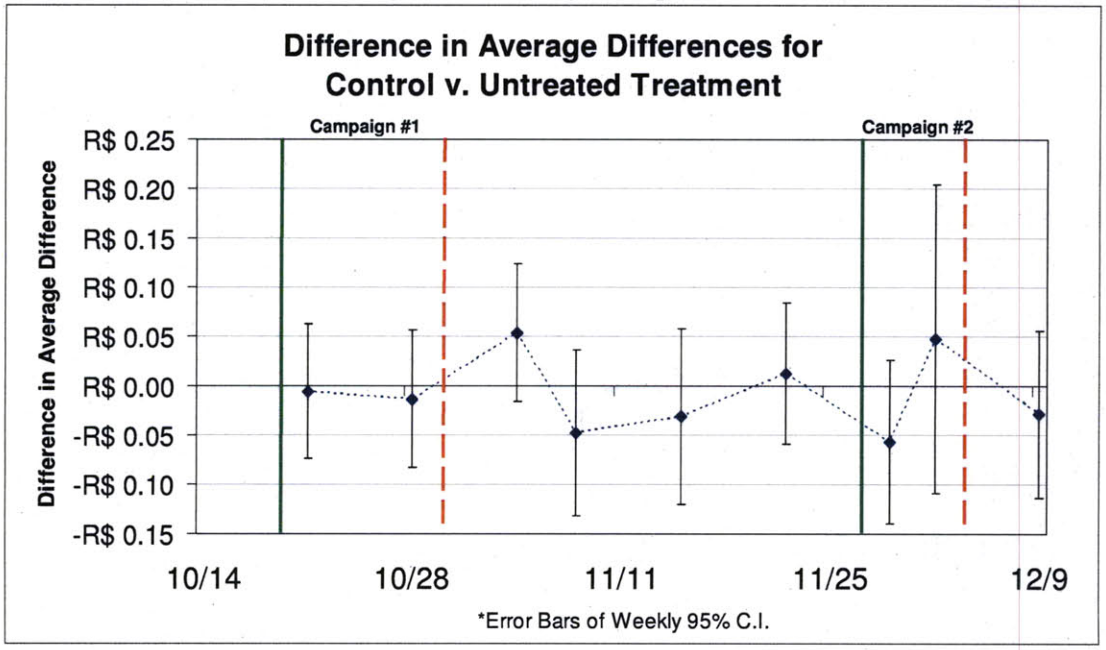


Table 4-6 - Results Summary for Both Campaigns

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

Figure 4-9 - Weekly DID Specification Test



*Error Bars of Weekly 95% C.I.

Table 4-7 - Offline/Online Treatment Effect Decomposition

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

Figure 4-10 - Nonparametric Estimate of the Treatment Effect by Ad Viewing Outcome

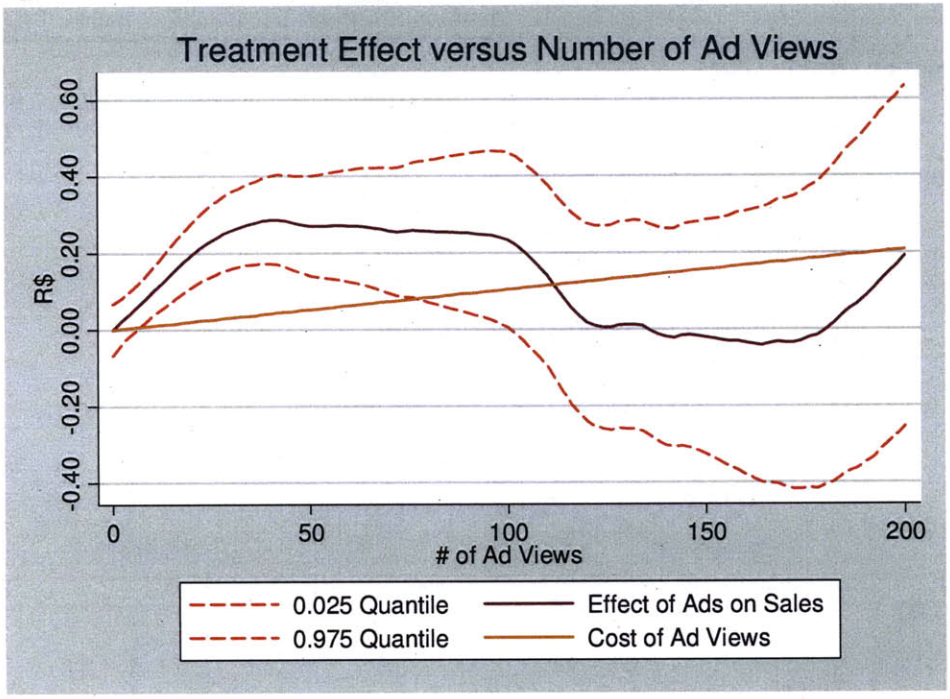
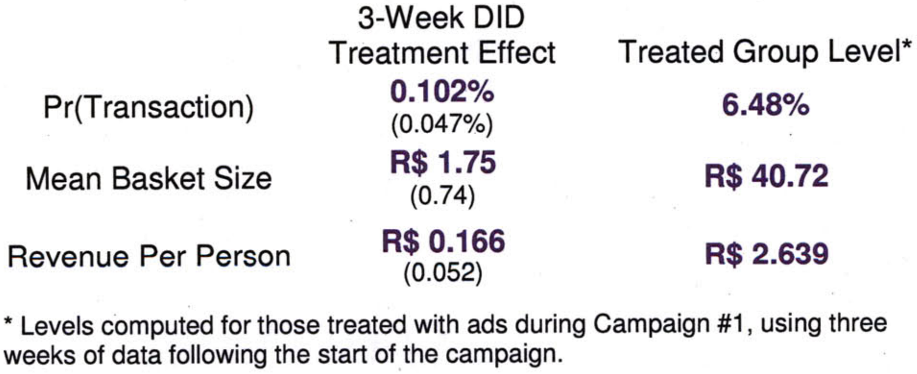


Table 4-8 - Decomposition of Treatment Effect into Basket Size and Frequency Effects



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

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

Chapter 5

Northern Exposure:

A Field Experiment Measuring Externalities between Search Advertisements

David H. Reiley, Sai-Ming Li, and Randall A. Lewis*

Abstract

"North" ads, or sponsored listings appearing just above the organic search results, generate the majority of clicks and revenues for search engines. In this paper, we ask whether the competing north ads exert externalities on each other. In particular, does increasing the number of rival north ads decrease the number of clicks I receive on my own north ad? We conduct a controlled experiment to investigate this question and find, to our surprise, that additional rival ads in the north tend to increase rather than decrease the click-through rate (CTR) of the top sponsored listing. We propose several possible explanations for this phenomenon, and point out directions for new theoretical models of sponsored search.

General Terms

Economics, Experimentation, Measurement

Keywords

Sponsored search, advertising, externalities, page placement

* Reiley, reiley@yahoo-inc.com; Li, eliot@yahoo-inc.com; and Lewis, ralewis@yahoo-inc.com. All authors can be contacted at Yahoo! Inc., 4401 Great America Parkway, Santa Clara, CA 95054. The authors would like to thank Kannan Achan, Erick Cantu-Paz, Rukmini Iyer, Ben Shahshahani, and Ajay Shekhawat for their advice and assistance in conducting the experiment. We thank Dan Ackerberg and Preston McAfee for helpful discussions.

I. Introduction

Online search engines respond to a query with a combination of two types of search listings: sponsored and organic. Search-engine algorithms produce the (unpaid) organic results, while advertisers pay for placement among the sponsored listings. The placement of a sponsored result depends on the competing advertisers' bids as well as the quality of each advertisement, where quality represents the propensity with which users feel inspired to click an ad. All else being equal, ads with higher bids appear higher up in the listings, as do ads with higher quality. Sponsored ads with sufficiently high bids and sufficiently high quality for a given query appear directly above the organic listings in what is known as the "north" position. Lower-ranked ads appear to the right of or below the organic listings in what are known as the "east" and "south" positions, respectively. This placement strategy aims to maximize search-engine revenue while delivering a high quality search experience to users.

From the search engine's point of view, there is a tradeoff in changing the number of north listings: more north listings can increase revenue by increasing the number of paid clicks, but may do so at the expense of user experience by displacing organic listings that might well be more relevant to the user. Search engines limit the number of north ads for this reason, using a variety of direct and indirect constraints on quality and expected revenue. For example, the top three search engines (Google, Yahoo!, Bing) never show more than four north ads on any query but quite frequently show zero or only one.

Advertisers are willing to bid higher in order to get better placement, and therefore more clicks, in the generalized second price auction (GSP) used by most search engines [1,2]. Knowing that additional clicks are valuable to the advertiser, we start this paper with the question: Would an advertiser be willing to pay for exclusive placement in the north?

One might reasonably assume that north ads would exert negative externalities on each other [3]. In particular, if I occupy the top advertising position and my ad appears in the north, then the presence of other ads below me in the north would likely take clicks away from me. If this is the case, advertisers would be willing to pay a premium for exclusive north placement, with the size of the premium determined by the number of additional clicks that exclusivity would generate.

Early published models of sponsored search [1,2] ignore the possibility of externalities between ads. [3] and [4] provide theoretical models for such click externalities, while [5] provides both theory and empirical evidence showing that the identity of a rival advertisement matters: the presence of a strong rival ad can cause me to receive fewer clicks than I would receive in the presence of a weak rival ad. By contrast, in this paper we look at variation in the *number* of north ads rather than at the identities of the rivals.

To get clean measurements of externalities, we perform a controlled field experiment, randomly varying the number of ads on each query. To our surprise, we find that rival north ads impose a *positive*, rather than negative, externality on existing north listings. In other words, the top north listing receives more clicks when additional sponsored listings appear below it. We suggest some possible explanations for this phenomenon, and point out directions for future research. We particularly note the importance of distinguishing north versus east ads, as well of understanding consumer substitution between organic and sponsored results.

The rest of the paper is organized as follows: section II describes the experimental method, section III describes the results of the experiment, section IV provides possible explanations of the observations, and section V summarizes our findings and provides suggestions on future work.

II. Experimental Method

This goal of the experiment is to measure how additional sponsored listings in the north impact the click-through rate (CTR) of existing north listings. In order to gather enough data to estimate the CTR for the same position across different numbers of north ads, we randomize the number of north ads for a representative sample of search queries on Yahoo! Search. We present results aggregated across a sample of many queries, users, and sponsored listings.

We choose to run a well-designed field experiment rather than pursuing the more common empirical technique of carefully examining existing observational data [5,6]. Experiments provide powerful insight by establishing causality rather than mere correlation, a topic known in econometrics as "the identification problem." [7,8] In the next paragraph, we provide a concrete example to illustrate an identification problem in this context, showing how observational data might lead to spurious conclusions. Because experiments generate exogenous variation and thereby solve the various econometric identification problems that plague observational studies, controlled field experiments have become increasingly popular in economics. In the past fifteen years, economists have successfully implemented field experiments on the topics of auction bidding [9,10,11] charitable fundraising [12,13], employee compensation [14,15], and economic development policies [16,17], to name just a few of the most prominent examples.

By running a controlled field experiment in this context, we create exogenously varying numbers of north ads on a representative set of queries. If we had instead relied on naturally occurring observational data, we might well have reached mistaken conclusions. For example, suppose our search engine always showed four north ads for "car rental" and always showed one north ad for "Taurus." Further suppose that the CTR for the top north ad on "car rental" is 20% while the CTR for the north ad on "Taurus" is 5%. If we looked at this data, we would find that the top north ad gets more clicks with three rival ads than with no rival ads, but this represents spurious correlation rather than causality. What we really want to know is how, holding the query constant, the propensity to click on a given ad depends on the number of rival north ads. Randomizing the number of north ads over a set of queries allows us to identify the causal effects of the number of ads.

We obtained permission to experiment on a randomly chosen 2% of all users of Yahoo! Search for a two-week period. The experiment manipulated the number of north ads for a randomly chosen 10% of the searches for these 2% of users. For each of these chosen search results, we randomly assigned the search request to one of five treatments where we displayed zero, one, two, three, or four north ads, with each of the five being equally likely. In all other respects, we presented these search-results pages exactly as they would have appeared in the absence of the experiment. For example, if a query would normally have displayed two north ads and we assigned it to have three north ads, the experiment would promote the top east ad to the third north position and any other east ads up one position, while leaving the organic results unchanged.

III. Experimental Results

Our primary variable of interest is the CTR on ads in position i for queries where there are a total of j north ads present, with $j \geq i$. With j north ads present, this click-through rate CTR_{ij} is defined to be the number of clicks c_{ij} on listings in position i divided by the number of search impressions N_{ij} displaying a listing in position i and j north ads:

$$CTR_{ij} = c_{ij} / N_{ij} .$$

To quantify our statistical estimation error, we make use of the well-known convenience that the binomial distribution converges to the normal distribution. We construct confidence intervals and standard errors using the usual estimator of the binomial distribution's variance,

$$\hat{\sigma}^2_{CTR_{ij}} = \frac{CTR_{ij} \cdot (1 - CTR_{ij})}{N_{ij}}.$$

Note that this formula depends on the standard assumption that all impressions are independent events. We also make this independence assumption when computing the estimated variance of the difference between two CTRs. This implies that the variance of their difference is equal to the sum of their variances. In practice, a single Yahoo! Search user sometimes performs many searches, so if users are heterogeneous in their click behavior, the click actions may be positively correlated across a given user's searches. This would increase the variance of our estimates of CTR, causing our standard errors to be slightly underestimated under the independence assumption. However, since a user with multiple searches would tend to be found in multiple treatments, any positive correlation would tend to produce minimal bias in our standard errors for the *difference* between two CTRs. We therefore expect any violations of the independence assumption to have little quantitative effect on our main results.

In the experiment, some searches had fewer than four ads available to display. For these searches, not all treatments were feasible—the server could not show more ads in the north position than there were ads available. We exclude these queries from our analysis in order to avoid selection bias. For example, for queries with only one ad available, suppose that the ad tends to be less relevant than average. Then this query would only get included in the analysis for zero or one north ads, but not two to four north ads. This would tend to bias the results towards finding positive externalities, because these less-clickable ads would only be included in results with less than two north ads. This would pull down the average CTR for the top position when j=1, but not when j≥2. Our chosen analysis enables us to compare apples to apples, studying the same population of ads in all treatments: all queries with at least four ads available. This excludes 25% of searches with at least two ads, leaving us with approximately 100,000 search observations per treatment.

A. North CTR

Figure 5-1 shows the CTR for each north position and number of north listings in our first experiment (October 2008). Surprisingly, at almost every rank the CTR *increases* with the number of listings shown in the north. The effect is especially pronounced in position one. We found this positive externality so unexpected in this first experiment that we replicated the experiment in order to validate the results. Figure 5-2 presents the results from the second experiment (June 2009), which involved an 18% larger sample than the first experiment. We limit the remaining analysis for this paper to the second experiment for clarity and brevity of exposition, choosing the second experiment because it contains more observations and more recorded variables than the first. All results for which we had data available in the first experiment are qualitatively identical to the first experiment, so our stated results actually understate the statistical power of our aggregate findings.

Figure 5-3 gives a more detailed view of the top-position north-ad results shown at the left of Figure 5-2. The first-north-ad CTRs have narrow 95% confidence intervals (hereafter all confidence intervals are 95% asymptotic confidence intervals). Upon comparing the CTRs, we find that displaying four north ads generates a statistically significantly higher CTR for the first-position ad (p=0.000) than for any of the alternatives of one, two, or three north ads. Strikingly, increasing the number of north ads tends to *increase* the CTR for the first north ad.

For positions two or three, we do not observe any statistically significant difference in CTR for sponsored listings when the number of north ads increases, as shown when moving to the right of the first cluster of bars in Figure 5-2. We suspect that this is primarily an issue of statistical power and hypothesize that with a larger sample a similar increasing trend would emerge for sponsored listings in positions two and three as the number of north listings increases. However, the short confidence intervals available let us safely conclude that the external effects are smaller than 5% of the baseline CTR without additional ads. By contrast, we estimate the positive externality on the first north position to be 2.2%, 4.7%, or 12.0%, when increasing from one north ad to two, three, or four north ads, respectively.

B. South and East Ad CTR

Figure 5-4 shows the CTR for each south and east position and number of north listings. The CTR declines with the number of north listings for almost every position, with the effect especially pronounced and statistically significant for the first two east and south positions. However, the data prevents us from making conclusive findings, especially in the lower positions.

We caution that, in contrast to the north-ad CTR results, these results do not have a clean causal interpretation. First, we note that the identities of the south and east ads vary to some extent with the number of north ads on the page. The set of ads displayed in a given position does not hold constant as we vary the number of north ads. For a given query, the available ads are ranked from first to twelfth prior to their placement on the page. In order, these ads are first placed into the available north positions (from zero to four) and then into the eight east (and south) positions until there are no more available slots. Thus, part of the result in Figure 5-4 may be due to selection effects. We see a decline in the CTR of east ads as the number of north ads increases, and it seems natural to interpret this as the additional north ads stealing clicks from the ads in the east positions. However, at least part of this decline in CTR may be due to lower ranked, less clickable ads occupying the east positions when more north ads are shown.

Similarly, many queries lack twelve eligible ads for placement. More competitive queries, attracting more advertisers, may tend to have higher CTRs. This is a reason why the CTR may fail to decline with the number of north ads in the later positions: with more north ads, the lower positions have fewer queries (i.e., a smaller denominator in the CTR calculation), and these queries tend to have more clickable ads on average. To examine this problem, we also computed CTRs based on ad ranking rather than position as seen in Figure 5-5. Ads in any given ranking monotonically receive more clicks when the number of north ads increases, moving them to a more preferred position.

One obvious implication of our north-ad results in Section III.A is that the search engine might want to consider the possibility of placing more than four north ads on certain queries. We provide one caution about that possible recommendation. Note that the south positions are bundled with the top two east ads on Yahoo! Search. This is important because the relevant CTR to the advertiser would then be the combined CTR. Note that when there are four north ads, the north ad CTRs are 12%, 6%, 4%, and 3%, respectively (Figure 5-2), while the fifth-position ad has a combined east-south CTR of 2.3% (Figure 5-4). Increasing the number of north ads beyond four could potentially result in more clicks to the first non-north position than to the last north position. This would have negative effects on the incentive compatibility and transparency of the auction mechanism for the advertiser, who expects higher bids to result in weakly higher numbers of clicks.

While Yahoo! Search bundles south listings with the top east listings, other major internet search engines do not. Google does not include any south ads in the results and, hence, avoids this concern. Bing bundles the two south positions with the top two north positions. Ask and AOL both omit east listings and bundle several north positions with the top south positions, below which additional ad listings are sometimes included. We highlight that the above concern regarding the south ads is specific to Yahoo! However, the externalities estimated for the north ads likely generalize to all the major internet search engines, which similarly include top-ranked ad units directly above the organic results in the north position.

C. Organic CTR

Figure 5-6 shows the CTR for each organic search result, for each number of north listings. The CTR at each organic position decreases with the number of north ads. This indicates that some of the clicks gained by the additional north listings are at the expense of organic clicks. These results have a clear causal interpretation, unlike those in the previous section, as the identities of the search results do not change with the number of north ads.

D. Overall CTR

We are also interested in the CTR for an entire section of the results page: north, east, south, or organic listings. We define the CTR for page section *k*, when there are *j* north ads present, to be

$$CTR_j^k = \frac{\sum_i c_{ij}^k}{N_j}$$

where c^k_{ij} denotes the number of clicks at position *i* in section *k* when *j* north ads are shown. Note that if a single search produces clicks on more than one listing, both clicks will be counted, so this CTR could in principle exceed 1. Similarly, the whole-page CTR is simply the sum of the section CTR over all four sections: organic, north, east, and south. Figure 5-7 shows the CTR for the whole search-results page, broken down by section. The CTR of the north section increases with the number of north listings, while the CTR of each other section declines. However, the CTR of the entire search-results page is slightly increasing with the number of north ads, as can be seen more clearly in Figure 5-8. Search engines, and publishers in general, prefer to increase advertising revenues with additional ad placements, so long as this does not significantly lower the utility of users. If we assume that all clicks are equally valuable to the user, then increasing the number of north ads from 0 to 4 does not appear to lower the value of the user experience.

However, it is possible that the additional clicks have lower utility to the user than existing ones. We next look at the fraction of searches by the number of clicks on links to better understand the impact of more north listings on user experience.

E. Number of clicks per search

Figure 5-9 shows the number of clicks per search under the varying number of north ads. To highlight the statistically significant differences, Figure 5-10 and Figure 5-11 show truncated bar charts of the fraction of searches with more than one click and zero clicks, respectively. The fraction of searches with no click does not vary systematically with the number of north listings. However, the fraction of searches with multiple clicks increases steadily with the number of north listings. Comparing Figure 5-10 and 5-11 shows that multiple-click behavior varies more widely (range ~2%) than zero-click behavior (range ~1%). Putting more sponsored listings in the north transforms a small number of one-click searches into multiple-click searches. Multiple clicks on the same search page is not necessarily an indication of bad user experience, because users may find more than one highly relevant result and want to explore them all. However, in this experiment, since sponsored listings are often less relevant than the organic listings they displace, the results suggest that increasing the number of north ads may cause users to have to work harder to find what they want.

IV. Explanations

We offer two hypotheses that might explain the experimental results. These hypotheses are not mutually exclusive; both may be relevant.

A. Substitution away from organic listings

Most theories of search advertising simplify the click-generation process in a way that would rule out the results obtained in this experiment. Early theories, such as [1,2] treated the consumer as a black box and made assumptions about CTRs that decline by position in the search listings. More recently, some theories have modeled consumer choice explicitly, but these simplify the choice problem by assuming away the organic search results [4,18]. The existing theoretical models cannot explain our results, which indicate (see especially Section III.C) that user click behavior depends on an interaction between sponsored listings and organic search results.

Existing models of human interaction with search results [4,5,18] make a reasonable top-down or "cascade" assumption that users, expecting the top results to be most relevant, start at the top of the page and scan through the search results sequentially until making a choice. Heuristics of this sort make sense for users, given the time cost of reading all the information on the page.

If users look at sponsored links in the context of the organic search results, then such heuristics could explain our results. Including an additional sponsored listing pushes the top organic results farther down the page, potentially making a sponsored result seem more relevant by comparison with what's shown directly below it. Such results can also be enhanced by psychological framing effects, as described in [19] and [20].

One possible consequence of this behavior is that, in the presence of additional north ads, users may find the first listing they click on does not fulfill their need, as they may have neglected the more relevant organic listings before their first click. In this case, users will have to go back to the search result page to find other listings. The result that the fraction of multiple-click searches increases with number of north listings supports this hypothesis.

B. *Comparative search behavior*

Another hypothesis is that when users are presented with multiple sponsored listings with similar messages, they tend to click on more than one such listing to see what each has to offer. In particular, some users may take multiple listings as a signal that the ads as a group provide something valuable, and this comparison-shopping motive makes them more likely to click on the first ad than they would if it were a singleton and more likely to click multiple ads.

In our assessment of positive externalities of the number of ads, we implicitly made the standard theoretical assumption that all clicks are equally valuable. If comparison-shopping behavior is the explanation for our results, then this assumption may be incorrect. An advertiser may get increasing numbers of clicks when the search engine increases the number of rival north ads, but these additional clicks may be less valuable because they are less likely to result in a conversion than the clicks obtained in the absence of the rival ads. Understanding this issue has important implications both for advertiser profits and for search-engine revenues, as bids depend on the advertiser's value of a click. Unfortunately, conversion data were unavailable to us in this experiment, so it remains an open question whether the rival north ads really exert positive externalities when both the number of clicks and the value of a click are taken into account.

V. Summary and Future Work

A. *Summary*

To better understand user click behavior in sponsored search, we have conducted an experiment to measure the externalities imposed by additional north listings on existing ones. Surprisingly, additional north listings impose a positive—not a negative—click externality on existing north listings. In other words, a given north listing has a higher CTR when more sponsored listings are shown below it. The increase in the CTR does not come at the expense of the additional north ads, but partially at the expense of the organic results. However, the total number of clicks registered on the page also increases, suggesting that increasing the number of north ads provides users an incentive to click on more links on the page. We have proposed two possible explanations for these results, and below we suggest additional research to explore them.

B. *Future research*

This experiment allowed us to learn several surprising results about the effect of increasing the number of north ads, particularly in the externalities imposed on existing ads by the inclusion of additional ads. This research suggests that a search engine might be able to create additional value by showing multiple north ads in some cases where currently only a single north ad appears. More research will better inform north-ad policy for search engines, as the best policy may depend on properties of the sponsored listings, the organic listings, the end user, and the value of additional clicks to the advertiser. We propose several directions for future experiments and analyses.

1. Examine heterogeneity by search keywords.

Different query keywords may exhibit greater or smaller amounts of the click externality discovered in this paper. Examining heterogeneity by keyword may yield some clues about the nature of the externality. For example, if the effect is larger for more commercial keywords, we conclude that the effect is due to the comparative-search hypothesis in Section IV.B. Unfortunately, this analysis is beyond the scope of our current dataset, which contains at most 200 searches for a given keyword. A new experiment could focus on a smaller set of perhaps 500 keywords, experimenting with 5,000 searches per keyword. A power calculation based on the existing data indicates that the click externality would be statistically significant for an individual keyword with a sample size of 5,000 searches. This calculation assumes that the externality has a similar magnitude to that found in the present experiment, with a 1.27% difference in the top-position CTR between one versus four north ads.

2. Examine heterogeneity by quality of the organic search listings.

Given our results indicating an interaction between organic results and north ads in user click behavior, we propose investigating how the externality varies with the quality of the organic listings. The hypothesis outlined in Section IV.A implies that higher-quality organic results, which tend to obtain more baseline clicks, are more likely to lose clicks to additional north ads. Unfortunately, our dataset does not contain good estimates of the quality of the organic search results, but in future research we intend to obtain such data, perhaps even using the randomized experiment to create additional variation in search-listing quality. Such research will provide insights into the ways that north-ad policy may affect both user experience and advertiser value.

3. Examine heterogeneity by browser display.

The total number of visible sponsored and organic listings on a search results page depends on users' window and font sizes. Users who can see more organic listings may behave differently from those who see fewer. If the north ads completely crowd out the organic results from the page display when first loaded, a user would be forced to scroll down to even find the first organic result. By tracking which results were immediately visible to users, we could test this source of heterogeneity in future experiments.

4. Measure the effect on the value of a click.

While users are clicking more when more ads are shown in the north, those clicks may have a lower conversion rate than if only one ad is shown in the north. If the additional clicks lead to at least proportionally more conversions, then we can safely conclude that increasing the number of north ads has a positive revenue externality for an advertiser, not merely a positive click externality. Obtaining conversion data in future experiments would help to settle this question.

VI. References

[1] Varian, H. R. 2007. Position auctions. Int. J. Ind. Organ. 25 (Dec. 2007) 1163-1178.

[2] Edelman, B., Ostrovsky, M., and Schwarz, M. 2007. Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords. Am. Econ. Rev. 97 (Mar. 2007) 242-259.

[3] Ghosh, A. and Mahdian, M. 2008. Externalities in online advertising. In Proceeding of the 17th international conference on World Wide Web (New York, NY, USA). WWW '08. ACM Press, New York, NY, USA, 161-168. DOI= http://doi.acm.org/10.1145/1367497.1367520

[4] Kempe, D. and Mahdian, M. 2008. A Cascade Model for Externalities in Sponsored Search. In Proceedings of the 4th International Workshop on Internet and Network Economics (Shanghai, China). WINE '08. Springer-Verlag, Berlin, Heidelberg, 585-596. DOI= http://dx.doi.org/10.1007/978-3-540-92185-1_65

[5] Gomes, R., Immorlica, N., and Markakis, E. 2009. Externalities in Keyword Auctions: An Empirical and Theoretical Assessment. In Proceedings of the 5th International Workshop on Internet and Network Economics (Rome, Italy). WINE '09. Springer-Verlag, Berlin, Heidelberg, 585-596. DOI= http://dx.doi.org/10.1007/978-3-642-10841-9_17

[6] Segal, I. and Jeziorski, P. 2009. What Makes Them Click: Empirical Analysis of Consumer Demand for Internet Search Advertising, Working Paper, Feb. 2009.

[7] John A. List and David H. Reiley, "Field Experiments." Forthcoming, The New Palgrave Dictionary of Economics, Steven N. Durlauf and Lawrence E. Blume, eds., Palgrave Macmillan Publishing: New York, NY, USA.

[8] Harrison, G. W. and List, J. A. 2004. Field Experiments. J. Econ. Lit. 42 (4) 1009-1055.

[9] Lucking-Reiley, D. H. 1999. Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet. Am. Econ. Rev. 89 (Dec. 1999) 1063-1080.

[10] List, J. A. and Reiley, D. H. 2000. Demand Reduction in Multi-Unit Auctions: Evidence from a Sportscard Field Experiment. Am. Econ. Rev. 90 (Sept. 2000) 961-972.

[11] Angrist, J. D. and Krueger, A. B. 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. J. Econ. Perspect. 15 (Fall 2001) 69-85.

[12] List, J. A. and Reiley, D. H. 2002. The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign. J. Polit. Econ. 110 (Feb. 2002) 215-233.

[13] Karlan, D. and List, J. A. 2007. Does Price Matter in Charitable Giving? Evidence from a Large-Scale Field Experiment. Am. Econ. Rev. 97 (Dec. 2007) 1774-1793.

[14] Bandiera, O., Barankay, I., and Rasul, I. 2005. Social Preferences and the Response to Incentives: Evidence from Personnel Data. Q. J. Econ. 120 (Aug. 2005) 917-962.

[15] Bertrand, M. and Mullainathan, S. 2004. Are Emily and Greg More Employable than Lakeesha and Jamal? A Field Experiment on Labor-Market Discrimination. Am. Econ. Rev. 94 (Sept 2004) 991-1013.

[16] Duflo, E., Kremer, M., and Robinson, J. 2009. Nudging Farmers to Use Fertilizer: Evidence from Kenya. Working paper. MIT. (Jan. 2009).

[17] Jensen, R. T. and Miller, N. 2008. Giffen Behavior and Subsistence Consumption. Am. Econ. Rev. 98 (Dec. 2008) 1553-77.

[18] Athey, S. and Ellison, G. 2009. Position Auctions with Consumer Search. NBER Working Paper Series no. 15253.

[19] Tversky, A. and Kahneman, D. 1981. The Framing of Decisions and the Psychology of Choice. Science 211 (Jan. 1981) 453-458.

[20] Schwartz, B. 2004. Paradox of Choice: Why More is Less. HarperCollins Publishers, Inc. New York, NY, USA. 126-128

VII. Figures

Figure 5-1. CTR of north sponsored listings appears to increase with the number of north ads in the first experiment.

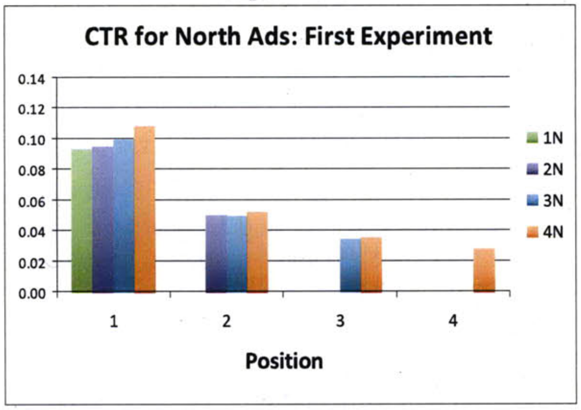


Figure 5-2. The replication verifies that the CTR of north sponsored listings increases with the number of north ads.

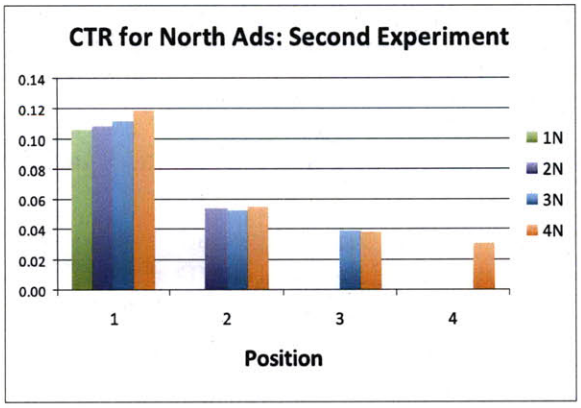


Figure 5-3. Truncated bar chart of position one north ad's CTR varying the total number of north ads.

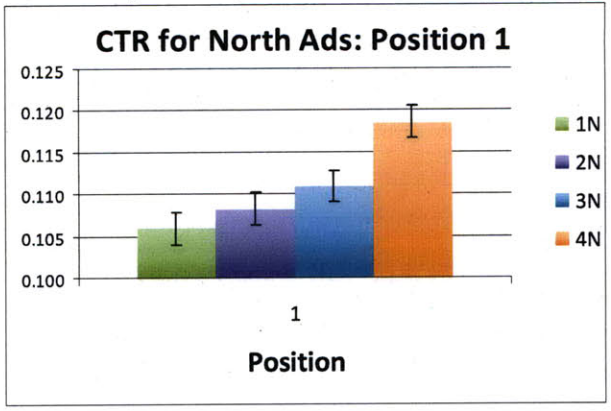


Figure 5-4. CTR of south and east ads declines with the number of north ads due to both ad quality and externalities.

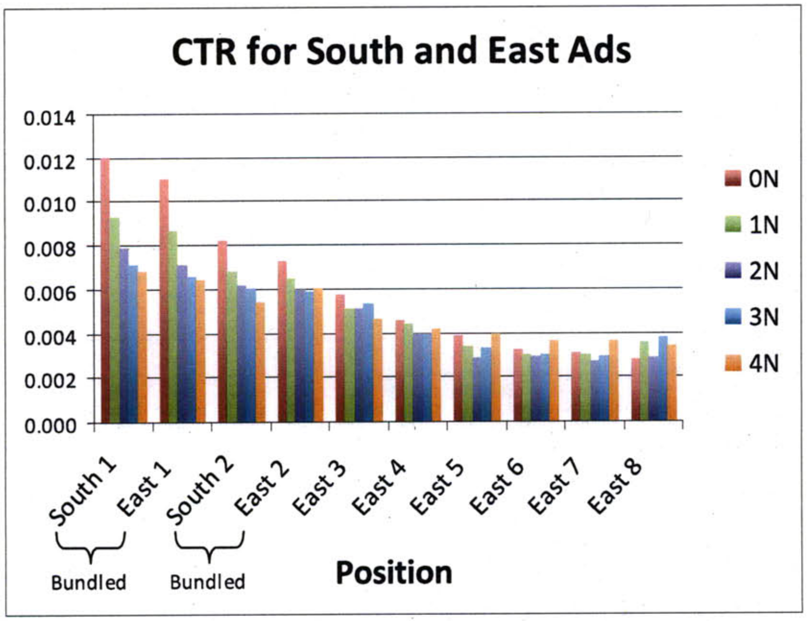


Figure 5-5. CTR for each ad rank increases with the number of north ads due to both position and externalities.

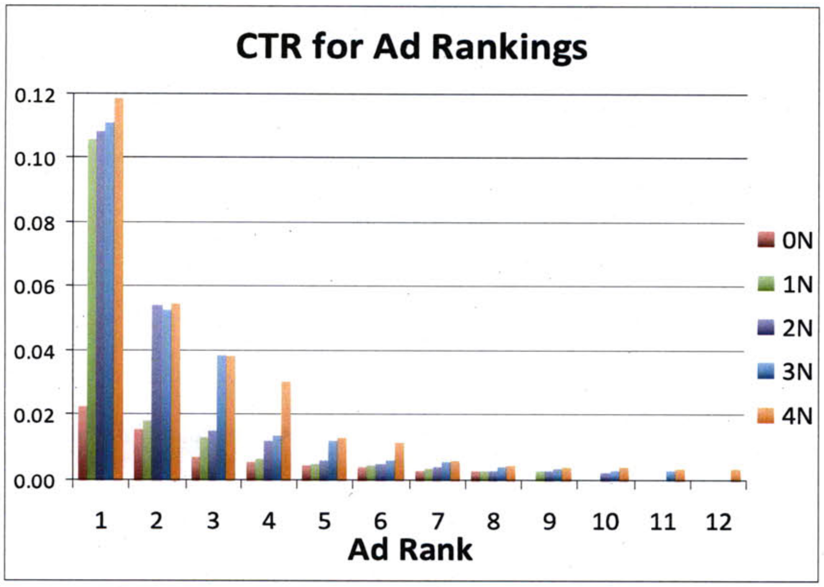


Figure 5-6. CTR of organic listings declines in the number of north ads for all organic search positions.

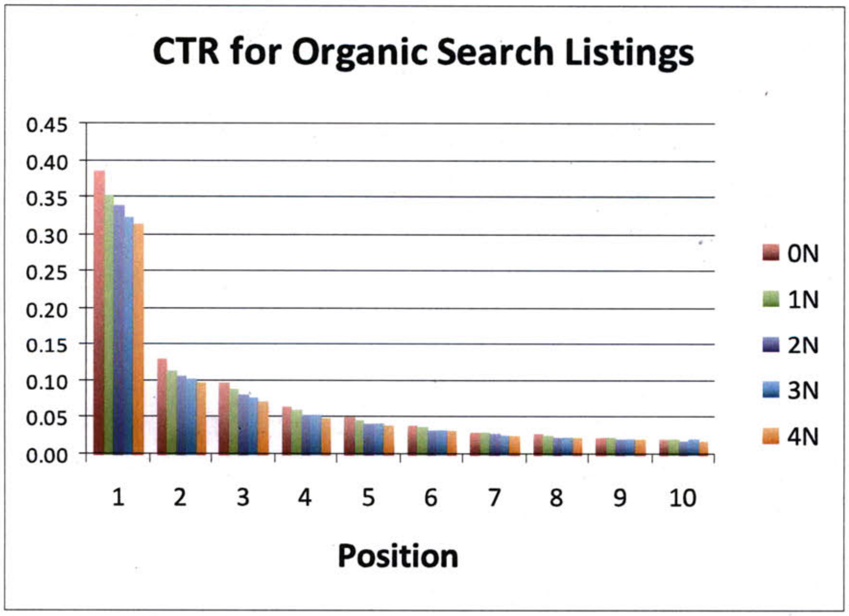


Figure 5-7. CTR by section shows an increasing number of north clicks at the expense of clicks in other sections.

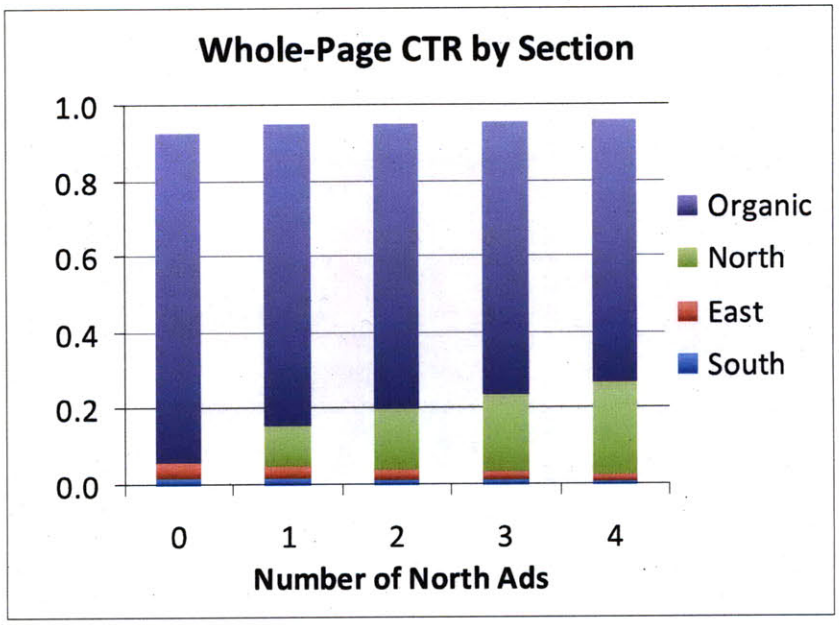


Figure 5-8. However, whole-page CTR per search still increases with the number of north ads.

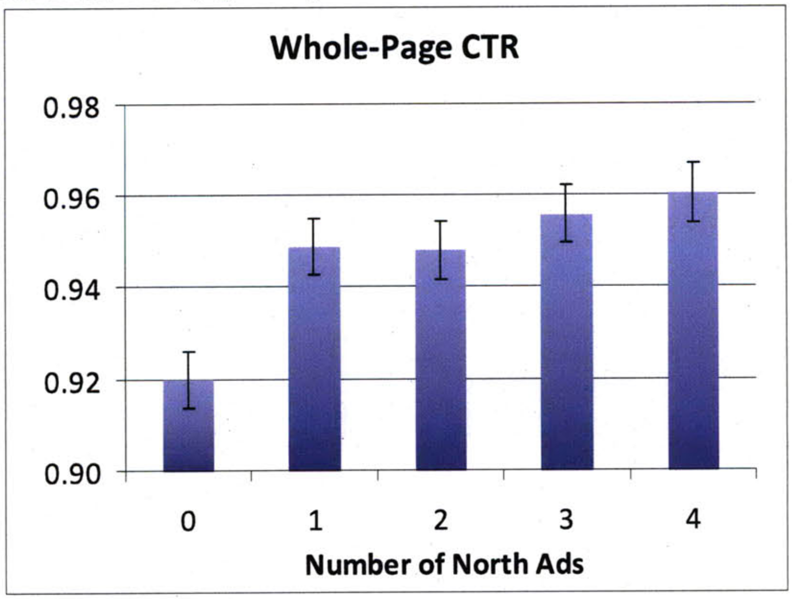


Figure 5-9. Fraction of searches divided into zero, one, and more than one click varies with the number of north ads.

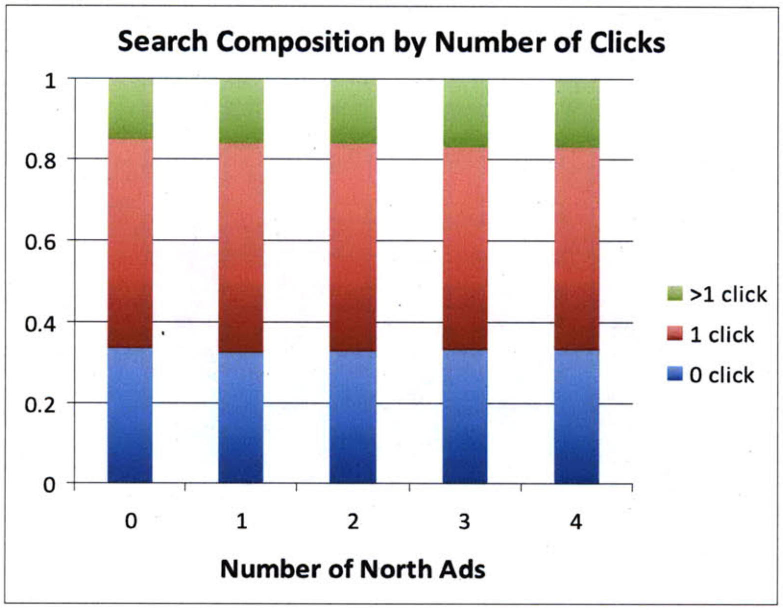


Figure 5-10. Fraction of searches with more than one click increases with the number of north ads.

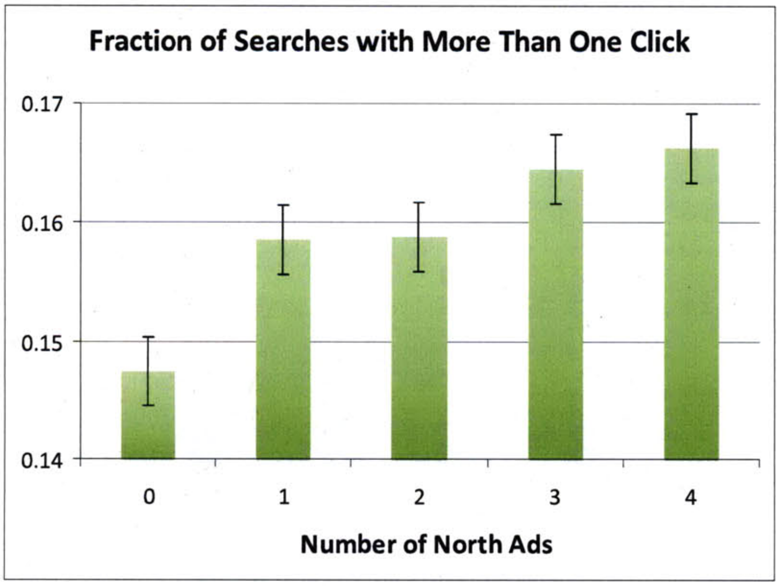
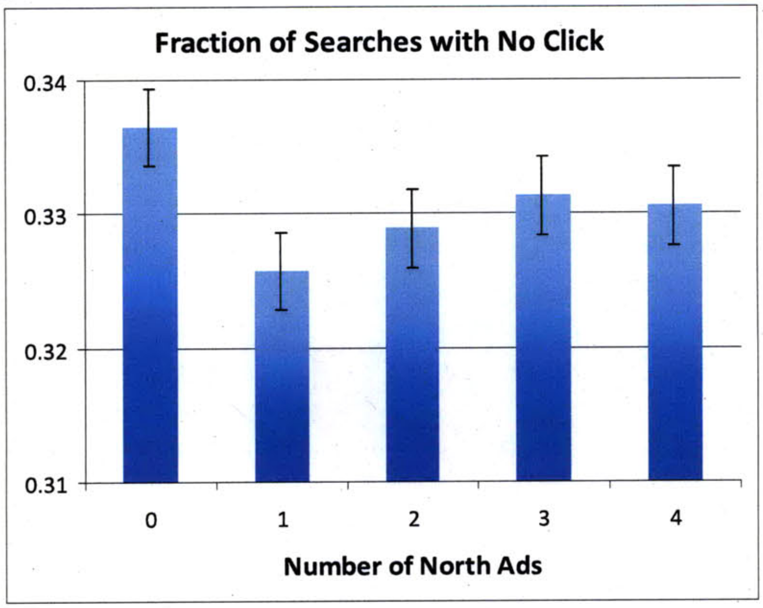


Figure 5-11. Fraction of searches with no click varies across different number of north ads.


Fraction of Searches with No Click
Number of North Ads