On the Near Impossibility of Measuring the Returns to Advertising*

Randall A. Lewis
Google, Inc.
ralewis@google.com

Justin M. Rao
Microsoft Research
justin.rao@microsoft.com

August 9, 2013

Abstract

Classical theories assume the firm has access to reliable signals to measure the causal impact of choice variables on profit. For advertising expenditure we show, using twenty-five online field experiments with major U.S. retailers and brokerages ($2.8 million expenditure), that this assumption typically does not hold. Evidence from the randomized trials is very weak because individual-level sales are incredibly volatile relative to the per-capita cost of a campaign—a "small" impact on a noisy dependent variable can generate positive returns. A calibrated statistical argument shows that the required sample size for an experiment to generate informative confidence intervals is typically in excess of ten million person-weeks. This also implies that selection bias unaccounted for by observational methods only needs to explain a tiny fraction of sales variation to severely bias observational estimates. We discuss how weak informational feedback has shaped the current marketplace and the impact technological advances moving forward.

Keywords: *advertising, field experiments, causal inference, electronic commerce, return on investment, information*
JEL Codes: *L10, M37, C93*

*Previous versions circulated under the name "On the Near Impossibility of Measuring Advertising Effectiveness." We especially thank David Reiley for his contributions to this work. He also came up with the title. Ned Augenblick, Arun Chandrasekhar, Garrett Johnson, Clara Lewis, R. Preston McAfee, Markus Möbius, Michael Schwarz, and Lars Lefgren gave us valuable feedback as well. We also thank attendees at Brigham Young University's Economics Seminar, the Becker Friedman Institute Advances with Field Experiments Conference, Wharton OPIM, and other venues where we have presented this work. We also thank countless engineers, sales people, and product managers at Yahoo, Inc. Much of this work was done when the authors were at Yahoo! Research, Santa Clara, CA.

1 Introduction

On a daily basis, the average American sees 25–45 minutes of television commercials, many billboards, and online ads (Kantar Media, 2008). Industry reports place annual advertising revenue in the U.S. in the range of $173 billion,[1] or about $500 per American per year. To break even, the universe of advertisers needs to net roughly $1.50 in profits per person per day; our educated guess is that this roughly corresponds to about $4–6 in incremental sales per person per day, or about $3,500–5,500 per household per year.

These back-of-the-envelope calculations demonstrate that the prices of advertising imply a large causal impact on household purchases in order for advertising to be a profitable venture on average, yet, perhaps surprisingly, the effects of advertising on consumer behavior are poorly understood. As a motivating example, the U.S. Department of Defense spent $72.3 million on NASCAR sponsorship to bolster recruiting. In 2012, an amendment to the Armed Services Appropriations Bill that would have eliminated this spending was narrowly defeated. It appears most legislators agreed that the relevant metric was return on investment—opponents argued that it was positive (many incoming recruits reported seeing sports marketing), while proponents argued it was negative. Since there were no reliable figures on the matter, the debate slipped into the rhetoric, exemplified in this quote from amendment co-sponsor Jack Kingston (R-Ga), "If someone is going to sign away five or six years of their life, it's going to take more than an ad on an automobile." Tens of millions of dollars had been spent, and seemingly no one could reliably state its causal impact on the intended goal.[2]

Consistent with this example, papers in the advertising effectiveness literature often use "Do ads have any effect?" as the key hypothesis to test. This is perhaps epitomized by the *first sentence* of an influential paper by Abraham et al. (1990), "Until recently, believing in the effectiveness of advertising and promotion

[1] This figure, while not perfect, is consistent with published market reports. We obtained it from http://www.plunkettresearch.com/ which aggregates a few reputable sources. In Appendix Figure 6.5, we use another data source, the Coen Structured Advertising Dataset, to plot advertising spending since World War I. During the entire period spending as a percent of GDP was fairly stable at 1.5–2%, lending further credence to the figure quoted in the text.

[2] In 1978, the Navy performed a series of experiments in which recruitment was heightened in certain geographic areas (Carroll et al. 1985). The impact of advertising is not well estimated, but the impact of recruiters was shown to be significantly positive.

was largely a matter of faith." An opening that might seem a bit peculiar, given that before it was written American firms had spent approximately $4.6 trillion promoting their products and services.[3]

In this paper we address the underlying puzzle: if so much money is being spent on advertising, how could it be possible that economic agents have such imprecise beliefs on the returns? It turns out that a key assumption of the classical theory of the firm, namely that firms have access to reliable signals mapping choice variables to profit, tends to fail in this domain. This assertion is based on our analysis of 25 large-scale digital advertising field experiments from well-known retailers and financial service firms partnering with a large web publisher. In total, the experiments accounted for $2.8 million in expenditure. We find that even when ad delivery and consumer purchases can be measured at the individual level, linked across purchasing domains, and randomized to ensure exogenous exposure, forming reliable estimates on the returns to advertising is exceedingly difficult even with millions or tens of million person-week observations. As an advertiser, the data are stacked against you.

The intuition for these results can be gleaned from the following observation: the effect of a single campaign should be "small" in equilibrium. Ads are relatively cheap (typically $<$ $0.01 per delivery) so only a small fraction of people need to be "converted" for a campaign to be profitable. Using detailed sales data from our partner firms, we show that matters are further complicated by the fact that the standard deviation of sales, on the individual level, is typically 10 times the mean over the duration of a typical campaign and evaluation window. The advertiser is trying to estimate a relatively subtle effect in an incredibly noisy economic environment. As an example, a 30-second television commercial during one of the best known (and expensive) advertising venues in the United States, the NFL Super Bowl, costs between 2.0–3.5 cents per viewer.[4] If a Super Bowl ad has a impact on profits of 7 cents per viewer, it is wildly profitable, while if it has an impact of 1 cent per viewer, it loses the company $1–2 million. The line between boom and bust

[3]This figure ($4.6 trillion) encompasses total ad spending from 1919 through 1990 (publication date) in the Coen Structured Advertising Dataset and is denominated in real 2005 US dollars.

[4]This figure is not perfectly precise, but definitely in the ballpark. See for instance: http://money.cnn.com/2007/01/03/news/funny/superbowl_ads/index.htm. A 30-second Super Bowl TV spot is priced at $2.5–4.0 million reaching an estimated audience of 110 million viewers according to Nielsen TV ratings.

is narrow, and with the sales noise we document, even a sample size of *100 million individuals* may not be adequate to distinguish between them.

To motivate our empirical analysis we develop a simple model of the firm's advertising problem. The key model parameter is return on investment (ROI)—the profits generated through the advertising as a percentage of the costs. In online advertising, intermediate metrics such as clicks have become popular in measuring advertising effectiveness.[5] Our focus, however, is on what the firm presumably cares about in the end, namely profits and our data sharing agreements allow us to sidestep intermediate metrics. We show that even using the gold-standard for measuring treatment effects, a fully randomized experiment, massive trials (typically in the single-digit millions of person-weeks) are required to reliably distinguish disparate hypotheses such as "the ad had no effect" (-100% ROI) from "the ad was profitable for the firm" (ROI>0%). Answering questions such as "was the ROI 15% or -5%," a large difference for your average investment decision, or "was the *annualized* ROI at least 5%," a reasonable question to calibrate against the cost of capital, typically require at least hundreds of millions of independent person-weeks—nearly impossible for a campaign of any realistic size. And while ROI measures accounting profits and losses, determining the profit maximizing level of ROI requires one to estimate the underlying profit function. We briefly discuss the (rather incredible) difficulties of this enterprise.

Instead of endorsing observational methods, the shortcomings of experiments actually serve to highlight their lurking biases. Ads are by design not delivered randomly because marketers target across time, people, and contexts. So while the true causal effect should be relatively small, selection effects are expected to be quite large. Consider a simple example: if an ad costs 0.5 cents per delivery, each viewer sees one ad, and the marginal profit per "conversion" is $30, then only 1 in 6,000 people need to be "converted" by the ad to break even. Suppose targeting individual has a 10% higher baseline purchase probability (a relatively small degree of targeting), then the selection effect is *600 times larger* than the causal effect of the ad. Imagine running a regression to explain sales per individual (in dollars) as a function of whether or not she saw advertising. Based on the empirical sales volatility we observe and the magnitude of the effect we are trying to estimate—a

[5]For a discussion of complications that can arise from using these metrics, see Lewis, Rao, and Reiley (2013)..

35 cent effect on a variable with a mean of $7 and a standard deviation of $75—the R^2 for a *highly profitable* campaign is on the order of 0.0000054.[6] To use successfully employ an observational method, we must be sure we have not omitted any control variables or misspecified the functional form to a degree that would generate an R^2 on the order of 0.000002 or more, otherwise estimates will be severely biased. This seems to be an impossible feat to accomplish, especially when selection effects are expected to be orders of magnitude larger than the true causal effect.

We now provide a bit more detail on the experiments and the results. The 25 distinct field experiments were part of advertising campaigns that each had more than 500,000 unique users, most over 1,000,000. To create exogenous variation in ad exposure, we randomly held out eligible users from receiving an advertiser's online display ad.[7] Sales tracking (both online and offline, through data sharing agreements) allows us to estimate the underlying variability and trends in sales.

The median standard error on ROI for the 25 campaigns is a staggering 51%. Supposing the ad was profitable (>0% ROI), 9 of the 25 experiments lacked sufficient power to reject -100% ROI, even though most of these experiments had over a million unique individuals. Continuing with this line of analysis, we look at the experiments' ability to evaluate a series of hypotheses. For each, we determine how many experiments had adequate power to reject the null when the alternative is true and for those that did not, we calculate how big the experiment would have to be, assuming an endless supply of independent consumers to add to the campaign, to become statistically reliable. Only 3 of the 25 campaigns could reliably distinguish between a wildly successful campaign (+50%) from one that broke even (0% ROI); the median campaign would have to be 9 times larger. In fact, retailers with relatively high sales volatility would need to run campaigns more than a 100 times larger to reliably evaluate these disparate hypotheses. There is heterogeneity on this dimension: 5 campaigns would have had adequate power had they been only 2.3 times larger. As we draw the alternative and null hypotheses towards more standard

[6] $R^2 = \frac{1}{4} \cdot \left(\frac{\$0.35}{\$75}\right)^2 = 0.0000054$.

[7] An example display ad is shown in Appendix Figure 6.4. Unlike search ads, these involve creatives that are larger and include images and potentially motion such as Adobe Flash animation. They are typically paid per impression, as opposed to per click, to incentivize producing a high quality creative. In search advertising, link-based ads are text based, so the problem is lessened significantly and further mitigated by using "clickability" in adjusting the effective bid. For a more detailed exposition on search advertising, see (Edelman et al., 2007).

tolerances for investment decisions, like +/- 10%, the heterogeneity vanishes—the estimation problem becomes nearly impossible for all firms. The median sales campaign would have to be *62 times larger* (mean 421x) to reliably distinguish between 10% and 0% ROI. For campaigns designed to acquire new account sign-ups at financial firms, the situation is even worse; the median campaign would have 1241 times as large, which provides a better analog to goods with all-or-nothing consumption profiles, such as automobiles.

In the discussion section our primary goal is to stress test the generalizability of our results. We first show that the firms we study are fairly representative of advertisers generally in terms of sales volatility, margins, and size. Our tongue-in-cheek "Super Bowl 'Impossibility' Theorem" shows that even a massive, idealized experiment can be relatively uninformative for many advertisers. We theoretically demonstrate that our results were not driven by the campaign windows we chose or the level of targeting of the campaigns under study, while recent empirical work from a major advertiser's research lab helps confirm our findings from an entirely different perspective (Blake et al., 2013).

Moving on the implications on the market as a whole, scarce information means that there is little "selective pressure" on advertising levels across firms. Consistent with this reasoning, we document otherwise similar firms in the same industries having vastly different levels of advertising spending. Informational scarcity also places an importance of reputation for advertising agencies and sets massive publishers off to an advantage because they can carve out a monopoly on reliable feedback for the returns to advertising.

2 A Simple Model of the Advertiser's Problem

In this section we formalize the problem of campaign evaluation. Our model is exceedingly simple, designed to capture only the core elements of measuring advertising returns. If measurement is very difficult in this simplified world, it follows that doing so in the real-world is harder still.

2.1 Model

Full-blown optimization of advertising would include, among other things, selecting the consumers to advertise to, measuring the advertising technology across various media, and determining how those technologies interact. Our focus here is on measuring the returns to an advertising campaign. We define a campaign as a set of advertisements delivered to a set of consumers through a single media over specified (and typically short) period of time. Ex-post evaluation asks the question, "Given a certain expenditure and delivery of ads, what is the rate of return on this investment (ROI)?" Clearly optimization and evaluation are different concepts with evaluation the easier to accomplish goal of creating feedback that can later be used to optimize.

We assume that a campaign uses one publishing channel with a single "creative" (all messaging content such as pictures, text, and audio). A campaign is defined by c, the cost per user. For a given publishing channel, c determines how many "impressions" each user sees. We assume the sales impact the campaign has a given consumer is defined by a continuous concave function of per-user expenditure,[8] $\beta(c)$. One can easily incorporate consumer heterogeneity with a mean-zero multiplicative parameter on this function and then integrate this parameter out to focus on the representative consumer. Let m be the gross margin of the firm, so the $m * \beta(c)$ gives gross profit per person. Net profit per-person subtracts cost $m * \beta(c) - c$ and ROI measures net profit as a percentage of cost $\frac{\beta(c)m-c}{c}$ (total returns equals to $N \cdot \beta(c)m - N \cdot c$). In our simple model the only choice variable is c, or "how much should I advertise to each consumer"—we take the campaign's target population as given.

Figure 1 graphically depicts the model: c^* gives the optimal spend level, and c_h (h for "high") gives the spend level where ROI is exactly 0%. Any point past c_h and the firm has negative earnings on the spend whereas any point to the left of c^* the firm is under-advertising. For points in (c^*, c^h), the firm is over-advertising because marginal return is negative but average return, or ROI, is still positive.

The model formalizes the estimation of the average per-person impact of a given campaign on consumer behavior. In reality, multiple creatives are used, the actual quantity of ads delivered per person is stochastic (because exposure depends on

[8]For supportive evidence of concavity see (Lewis, 2010). This assumption could be weakened to "concave in the region of current spending," which essentially just says that the returns to advertising are not infinite and the firm is not in a convex region.

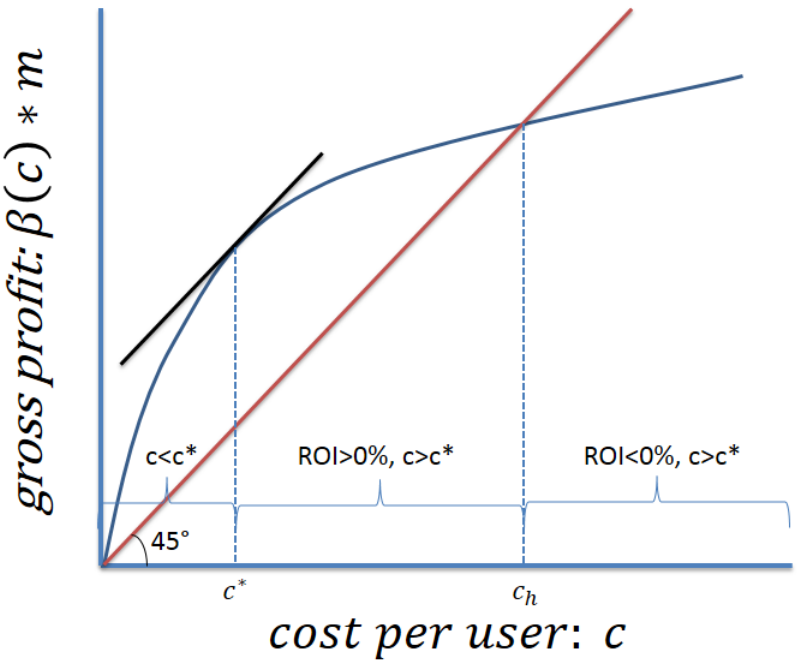


Figure 1: Graphical depiction of the advertiser's problem.

user activity), and β should include as arguments the following non-exhaustive list: tv/billboard/etc. spending in campaign period, consumer preferences, time of year, and time of day of the advertisement. The evaluation framework is motivated by the fact that the "campaign" is an important operational unit in marketing. A Google Scholar search of the exact phrase "advertising campaign" returned 48,691 unique research documents. This is confirmed by our personal experience, for campaigns of reasonable size firms typically evaluate performance at the campaign level, often to evaluate whether they should re-up the purchase order on the campaign.

2.2 Measuring the Returns to a Campaign

We first will start out with a high-level view of the inference challenges facing an advertiser by providing of overview of the market and calibrating the model using

median figures from the experiments. On the cost side, online display ad campaigns that deliver a few ads per day per targeted individual cost about 1–2 cents per person per day (this puts costs in the same ballpark as running one 30-second TV ad per person per day) and typically run for about two weeks, cumulating in a cost between 15 and 40 cents per person. Given the total volume of ads a consumer sees across all media, even an intense campaign only captures about 2% of a targeted person's advertising "attention"[9] during the campaign window.

We now need to quantify sales volatility. Sales volatility has three components: the average magnitude (mean sales), heterogeneity (variance of per-person means), and rarity of purchases (stochasticity in purchasing). For the large retailers and financial service firms in our study, the mean weekly sales per-person varied considerably across firms, as does the standard deviation in sales. However, we find that the ratio of the standard deviation to the mean (the coefficient of variation of the mean) is typically around 10 for the retail firms—customers buy goods relatively infrequently, but when they do, the purchases tend to be quite large relative to the mean.[10] Sales volatility tends to be higher for financial service firms, because people either sign-up and become lucrative long-term customers or they don't use the service at all.

In the econometric model let y_i be sales for individual i. Since we are assuming, for simplicity, that each affected individual saw the same value of advertising for a given campaign, we will use an indicator variable x_i to quantify ad exposure. $\hat{\beta(c)}$ gives our estimate of the sales impact $\beta(c)$ for a campaign of cost-per-user c. Standard econometric techniques estimate this value using the difference between the exposed (E) and unexposed (U) groups. In an experiment, exposure is exogenous. In an observational study, one would also condition on covariates W and a specific functional form, which could include individual fixed effects, and the following nota-

[9] Ads are typically sold by delivered impressions, but this does not necessarily mean a person noticed them. Indeed, one of the justifications for targeting is that a relevant ad is more likely to be noticed and read or watched. In reality, it is possible for a campaign to get 100% of a consumer's attention (he or she pays attention to that ad and ignores all others) or 0% (it is totally ignored) or any value in between.

[10] An extreme example of this feature is automobiles (which we discuss later) where the sales impact is either a number ranging in the tens of thousands of dollars, or more likely, given the infrequency of car purchases, it is \$0. Homogeneous food stuffs have more stable expenditure, but their very homogeneity likely reduces own-firm returns to and equilibrium levels of advertising within industry as a result of positive advertising spillovers to competitor firms (Kaiser, 2005).

tion would use $y|W$. All the following results go through with the usual "conditional upon" caveat.

For the case of a fully randomized experiment, our estimation equation is simply:

$$y_i = \beta x_i + \epsilon_i \tag{1}$$

We suppress c in the notation because a given campaign has a fixed size per user. The average sales impact estimate, $\hat{\beta}$, can be converted to ROI by multiplying by the gross margin to get the gross profit impact, subtracting per-person cost, and then dividing by cost to get the percentage return.

Below we use standard notation to represent the sample means and variances of the sales of the exposed and unexposed groups, the difference in means between those groups, and the estimated standard error of that difference in means. Without loss of generality we assume that the exposed and unexposed samples are the same size ($N_E = N_U = N$) and have equal variances ($\sigma_E = \sigma_U = \sigma$), which is the best-case scenario from a design perspective.

$$\bar{y}_E \equiv \frac{1}{N_E}\sum_{i\in E} y_i,\ \bar{y}_U \equiv \frac{1}{N_U}\sum_{i\in U} y_i \tag{2}$$

$$\hat{\sigma}_E^2 \equiv \frac{1}{N_E - 1}\sum_{i\in E}(y_i - \bar{y}_E)^2,\ \hat{\sigma}_U^2 \equiv \frac{1}{N_U - 1}\sum_{i\in U}(y_i - \bar{y}_U)^2 \tag{3}$$

$$\Delta\bar{y} \equiv \bar{y}_E - \bar{y}_U \tag{4}$$

$$\hat{\sigma}_{\Delta\bar{y}} \equiv \sqrt{\frac{\hat{\sigma}_E^2}{N_E} + \frac{\hat{\sigma}_U^2}{N_U}} = \sqrt{\frac{2}{N}} \cdot \hat{\sigma} \tag{5}$$

We focus on two familiar econometric statistics. The first is the R^2 of the regression of y on x, which gives the fraction of the variance in sales attributed to the campaign (or, in the model with covariates, the partial R^2 after first conditioning on covariates in a first stage regression—for a nice explanation of how this works, see Lovell, 2008):

$$R^2 = \frac{\sum_{i\in U}\left(\bar{y}_U - \bar{y}\right)^2 + \sum_{i\in E}\left(\bar{y}_E - \bar{y}\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2} = \frac{2N\left(\frac{1}{2}\Delta\bar{y}\right)^2}{2N\hat{\sigma}^2} = \frac{1}{4}\left(\frac{\Delta\bar{y}}{\hat{\sigma}}\right)^2. \tag{6}$$

Second is the t-statistic for testing the hypothesis that the advertising had no impact ($\beta = 0$):

$$t_{\Delta\bar{y}} = \frac{\Delta\bar{y}}{\hat{\sigma}_{\Delta\bar{y}}} = \sqrt{\frac{N}{2}}\left(\frac{\Delta\bar{y}}{\hat{\sigma}}\right). \tag{7}$$

In both cases, we have related a standard regression statistic to the ratio between the average impact on sales ($\Delta\bar{y}$) and the standard deviation of sales (σ)—we will call this the impact-to-standard-deviation ratio. It is also referred to as *Cohen's d* or, more generally, the signal-to-noise ratio.

We calibrate the test statistics using average values from 19 experiments run with large U.S. retailers in partnership with Yahoo! (the remaining 6 experiments were for account sign-ups for financial firms, making it harder to determine sales in dollars). For ease of exposition, we will discuss the hypothetical case as if it were a single, actual experiment. This representative campaign costs \$0.14 per customer, which amounts to delivering 20–100 display ads at a price of \$1-\$5 CPM,[11] and the gross margin is assumed to be 50%.[12] Mean sales per-person for the period under study is \$7 and the standard deviation is \$75.

We will suppose the ROI goal was 25%, meaning the goal was to generate a \$0.35 sales impact per person, yielding gross profits of \$0.175 per person as compared to costs of \$0.14. A \$0.35 per-person impact on sales corresponds to a 5% increase in sales during the two weeks of the campaign. The estimation challenge facing the advertiser is to detect this \$0.35 difference in sales between the treatment and control groups amid the noise of a \$75 standard deviation in sales. The impact-to-standard-deviation ratio is only 0.0047. From our derivation above, this implies an R^2 of:

$$R^2 = \frac{1}{4} \cdot \left(\frac{\$0.35}{\$75}\right)^2 = 0.0000054. \tag{8}$$

Even a successful campaign with a *large* ROI has R^2 of only *0.0000054*, meaning we'll need a very large N to reliably distinguish it from 0, let alone give a precise confidence interval. Suppose we had 2 million unique users evenly split between test and control in a fully randomized experiment. With a true ROI of 25% and an impact-to-standard-deviation ratio of 0.0047, the expected t-statistic for a com-

[11] CPM is the standard for impression-based pricing for online display advertising. It stands for "cost per mille" or "cost per thousand."

[12] We base this assumption on our conversations with retailers, our knowledge of the industry and SEC filings.

parison to -100% ROI (zero causal impact) is 3.30, using the above formula. This corresponds to a test with power of about 95% at the 10% (5% one-sided) significance level because the approximately normally distributed t-statistic should be less than the critical value of 1.65 about 5% of the time (corresponding to the cases where we cannot reject the null). With 200,000 unique users, the expected t-statistic is 1.04, indicating an experiment of this size is hopelessly underpowered: under the alternative hypothesis of a healthy 25% ROI, we fail to reject the null that the ad had no causal impact 74% of the time.[13]

The minuscule R^2 for the treatment variable in our representative randomized trial has serious implications for observational studies, such as regression with controls, difference-in-differences, and propensity score matching. An omitted variable, misspecified functional form, or slight amount of correlation between browsing behavior and sales behavior generating R^2 on the order of 0.0001 is a *full order of magnitude* larger than the true treatment effect. Meaning a very small amount of endogeneity would *severely bias* estimates of advertising effectiveness. Compare this to a classic economic example such as wage/schooling regressions, in which the endogeneity is roughly 1/8 the treatment effect (Card, 1999). It is always important to ask, "What is the partial R^2 of the treatment variable?" If it is very small, as in the case of advertising effectiveness, clean identification becomes paramount. As we showed in the introduction, a minimal level of targeting that results in the exposed group having a 10% higher baseline purchase rate can lead to an exposed-unexposed difference of about 600 times the true treatment effect. Unless this difference is controlled for with near *perfect* precision, an observational model's estimates have a large bias in expectation.

In showing the biases in observational methods, are we arguing against a straw man? Not so, according to a recent article in the *Harvard Business Review*. The following quotation is from the president of comScore, a large data-provider for web publishing and e-commerce:

> 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

[13]Note that when a low powered test does, in fact, correctly reject the null, the point estimates conditional on rejecting will be significantly larger than the alternatively hypothesized ROI. That is, when one rejects the null, the residual on the effect size is positive in expectation.

> viewed the ad, then compare online purchases made by those who have and those who have not seen it.
>
> M. Abraham, 2008.

The author used this methodology to report a 300% improvement in outcomes for the exposed group, which seems surprisingly high as it implies that advertising prices should be at least an order of magnitude higher than current levels.

3 Analysis of the 25 Field Experiments

In this section we delve deeper into the 25 experiments in the study. Due to confidentiality agreements, we cannot reveal the identity of the advertisers. We can say they are large firms that are most likely familiar to American readers.

3.1 Summary Statistics and Overview

Table 1 gives an overview of 25 display advertising experiments/campaigns. Some of the experiments are taken from past work out of Yahoo! Labs: Lewis and Reiley (2010); Lewis and Schreiner (2010); Johnson, Lewis, and Reiley (2011); and Lewis, Rao, and Reiley (2011).[14] We highlight the most important figures and present summary statistics. We employ a naming convention using the vertical sector of the advertiser in lieu of the actual firm names. The firms in Panel 1 are retailers, such as large department stores; in Panel 2 they are financial service firms.

Columns 1–3 of Table 1 give basic descriptors of the experiment. Columns 4–7 outline the outcome measures. Sales is the key dependent measure for the firms in Panel 1, which is shown in Column 4 along with the unit of observation ("3" indicates daily observation, "4" indicates weekly). In Panel 2, the dependent measure is new account sign-ups. Column 7 gives the control variables we have to reduce noise in the experimental estimates. The experiments ranged from 2 to 135 days (Column 8), with a median of 14 days, which is typical of display campaigns. Column 9 shows the campaign cost varied from relatively small ($9,964) to quite large ($612,693). The mean was $114,083; the median was $75,000. Overall, the campaigns represent

[14] We express gratitude to the other authors and encourage readers to examine these papers in more detail.

Table 1: Overview of the 25 Advertising Field Experiments

Retailers: In-Store + Online Sales*

			Estimation Strategies Employed**						**Campaign Level Summary**				**Per Customer**	
Adv	Year	#	Y	X	Y&X	W	Days	Cost	Assignment		Exposed		Avg. Sales	σ sales
									Test	Control	Test	Control	(Control)	
R 1	2007	1	1,4	1	-	1,2,3	14	$128,750	1,257,756	300,000	814,052	-	$9.49	$94.28
R 1	2007	2	1,4	1	-	1,2,3	10	$40,234	1,257,756	300,000	686,878	-	$10.50	$111.15
R 1	2007	3	1,4	1	-	1,2,3	10	$68,398	1,257,756	300,000	801,174	-	$4.86	$69.98
R 1	2008	1–6	1,4	1,2,3	-	1,2,3	105	$260,000	957,706	300,000	764,235	238,904	$125.74	$490.28
R 1	2010	1	1,4	1,2	-	1,2,3,4	7	$81,433	2,535,491	300,000	1,159,100	-	$11.47	$111.37
R 1	2010	2–3	1,3,4	1,2,3,4	1	1,2,3	14	$150,000	2,175,855	1,087,924	1,212,042	604,789	$17.62	$132.15
R 2	2009	1a	1,5	1	-	-	35	$191,750	3,145,790	3,146,420	2,229,959	-	$30.77	$147.37
R 2	2009	1b	1,5	1	-	-	35	$191,750	3,146,347	3,146,420	2,258,672	-	$30.77	$147.37
R 2	2009	1c	1,5	1	-	-	35	$191,750	3,145,996	3,146,420	2,245,196	-	$30.77	$147.37
R 3	2010	1	1,3,4	1,2,3	1	1,3	3	$9,964	281,802	161,163	281,802	161,163	$1.27	$18.46
R 3	2010	2	1,3,4	1,2	1	1,3	4	$16,549	483,015	277,751	424,380	-	$1.08	$14.73
R 3	2010	3	1,3,4	1,2,3	1	1,3	2	$25,571	292,459	169,024	292,459	169,024	$1.89	$18.89
R 3	2010	4	1,3,4	1,2,3	1	1,3	3	$18,234	311,566	179,709	311,566	179,709	$1.29	$16.27
R 3	2010	5	1,3,4	1,2	1	1,3	3	$18,042	259,903	452,983	259,903	-	$1.75	$18.60
R 3	2010	6	1,3,4	1,2,3	1	1,3	4	$27,342	355,474	204,034	355,474	204,034	$2.64	$21.60
R 3	2010	7	1,3,4	1,2,3	1	1,3	2	$33,840	314,318	182,223	314,318	182,223	$0.59	$9.77
R 4	2010	1	1,3,4	1,2	1	1	18	$90,000	1,075,828	1,075,827	693,459	-	$0.56	$12.65
R 5	2010	1	1,5	1,2	-	1,3	41	$180,000	2,321,606	244,432	1,583,991	-	$54.77	$170.41
R 5	2011	1	1,3,4	1,2	1	1,3	32	$180,000	600,058	3,555,971	457,968	-	$8.48	$70.20

Financial Services: New Accounts Online Only***

			Estimation Strategies Employed**						**Campaign Level Summary**				**Per Customer**	
Adv	Year	#	Y	X	Y&X	W	Days	Cost	Assignment		Exposed	New	Pr New	SE New
									Test	Control		Accts	(Test)	Acct
F 1	2008	1a	2,5	1,2,4	-	3	42	$50,000	12% of Y!	52% of Y!	794,332	867	0.0011	0.0330
F 1	2008	1b	2,5	1,2,4	-	3	42	$50,000	12% of Y!	52% of Y!	748,730	762	0.0010	0.0319
F 1	2008	1c	2,5	1,2,4	-	3	42	$75,000	12% of Y!	52% of Y!	1,080,250	1,254	0.0012	0.0341
F 1	2008	1d	2,5	1,2,4	-	3	42	$75,000	12% of Y!	52% of Y!	1,101,638	1,304	0.0012	0.0344
F 2	2009	1	2,3	1,2,3,4	1,2	3	42	$612,693	90% of Y!	10% of Y!	17943572	10,263	0.0006	0.0239
F 2	2011	1	2,5	1,2	-	4	36	$85,942	8,125,910	8,125,909	793,042	1090	0.0014	0.0331

* These retailers do a supermajority of sales via their brick & mortar stores.

** Estimation strategies employed to obtain the standard errors of the ad impact between the test and control groups follow:
"Y" 1:Sales, 2:Sign-ups, 3:Daily, 4:Weekly, 5:Total Campaign Window;
"X" 1:Randomized Control, 2:Active on Y! Network or site where ads were shown, 3:Placebo Campaign for Control Group, 4:Multiple Treatments;
"Y&X" 1: Sales filtered post first exposure or first page view, 2:Outcome filtered based on post-exposure time window);
"W" 1:Lagged sales, 2:Demographics, 3:Online behaviors.

*** These financial services advertisers do a supermajority of their business online.

over $2.8 million in expenditure. The median campaign reached over one million individuals, and all campaigns had hundreds of thousands of individuals in both test and control cells (Columns 9–11).

The second-to-last column shows that the average sales per customer varied widely across the firms. This is driven by both the popularity of the retailer and the targeting level of the campaign (a more targeted campaign often leads to higher baseline sales in the sample). Median sales per person is $8.48 for the test period. The final column gives the standard deviation of sales on an individual level. The median campaign had a standard deviation 9.83 times the mean. We plot the distribution of standard-deviation-to-mean ratio against campaign duration in Figure 2. This ratio exceeds 7 for 23 of the 25 experiments. Longer campaigns tend to have a lower ratio, which is due to sufficient independence in sales across weeks.[15] While a longer campaign with the same sample size effectively has more data, these additional data will only make inference easier if the spending per person per week is not diluted (see section 4.1.3); otherwise, the effect size is expected to fall, making it harder to detect.

3.2 Estimating ROI

In Table 2, we take a detailed look at estimating ROI. Column 3 gives the standard error associated with the estimate of β, the test-control sales difference as defined by the model (in dollars for Panel 1, in account sign-ups for Panel 2). We condition on the control variables outlined in Column 7 of Table 1 in order to get the a precise an estimate as possible. In Column 4, we give the implied radius (+/- window) of the 95% confidence interval for the sales impact, in percentage terms—the median radius is 5.5%. Column 5 gives the per-person advertising spend, which can be compared to the standard error of the treatment effect given in Column 3 to capture how the magnitude of statistical uncertainty relates to the expenditure. In Column 7 we translate the sales impact standard errors to ROI in percentage terms using our estimates of gross margins (Column 6, which are based on SEC filings). For the financial firms we convert a customer acquisition into a dollar value using a figure

[15] If sales are, in fact, independent across weeks, we would expect the coefficient of variation to follow $\frac{\sqrt{T}\cdot\sigma_{weekly}}{T\cdot\mu}$. However, over long horizons (i.e., quarters or years), individual-level sales are correlated, which also makes past sales a useful control variable when evaluating longer campaigns.

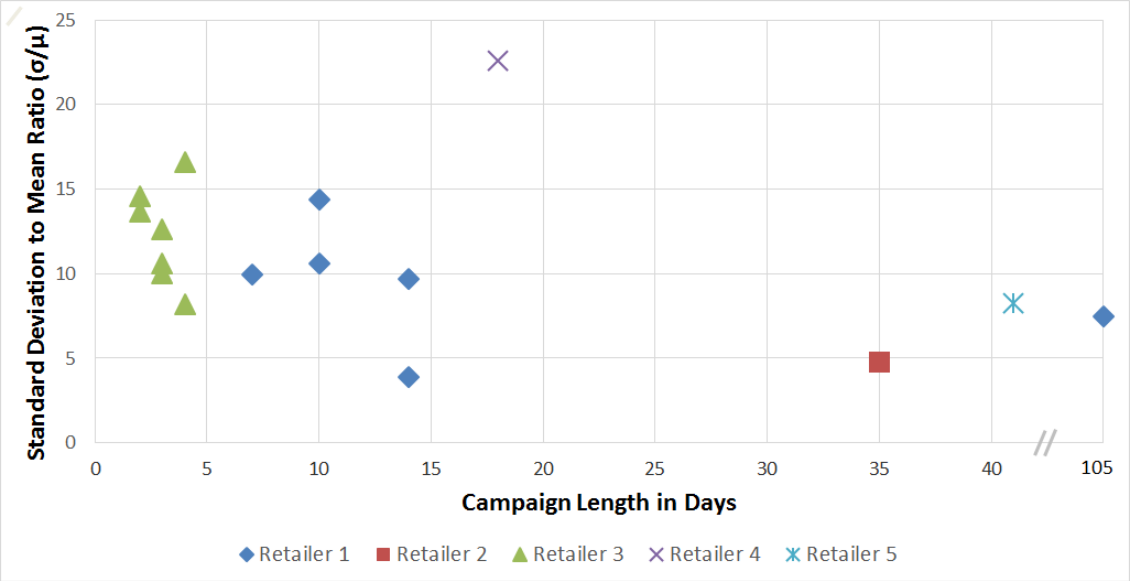


Figure 2: Relationship between sales volatility, as given by the coefficient of variation, $\frac{\sigma}{\mu}$, and campaign length in days.

provided for us by the advertisers.[16] The median standard error for ROI is 26.1%, it follows that the median confidence interval is about *100 percentage points wide*. The mean standard error is higher still at 61.8%, implying a confidence interval that is too wide to be of much practical use. Even with relatively large randomized experiments, estimating ROI is far from a precise enterprise.

In Figure 3 we plot the standard error of the ROI estimate against the per capita campaign cost. Each line represents a different advertiser. Two important features are immediately apparent. First, there is significant heterogeneity across firms. Retailer 1 and the financial firms had the highest statistical uncertainty in the ROI estimate. Financial firms operate in an all-or-nothing world—someone either signs up for an account (and likely becomes a lucrative long-term customer) or does not. Retailer 1 simply has a higher standard deviation of sales as compared to the other retailers. Second, estimation tends to get more precise as the per-person spend increases. The curves are downward sloping with the exception of a single point. This is exactly what we would expect. For a given firm, a more expensive

[16]This figure is presumably their estimate of the discounted net present value of a new customer acquired through online advertising.

Table 2: Statistical Precision and Power Calculations for the 25 Advertising Field Experiments

In-Store + Online Sales

		Key Statistical Properties of Campaign					HARD: Ads did anything? H0: ROI=-100% Ha: ROI=0%		HARDER: Worth money? H0: ROI=0% Ha: ROI=50%		HARDEST: To optimize? H0: ROI=0% Ha: ROI=10%		CRAZY: Profits maximized? H0: ROI=0% Ha: ROI=5%	
Adv	#	SE β Sales	Radius 95% CI % Sales	Spend Per Exposed	Margin	SE ROI	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3
R 1	1	$ 0.193	4.0%	$0.16	50%	61%	1.64	3.3x	0.82	13.4x	0.16	335x	0.08	1338x
R 1	2	$ 0.226	4.2%	$0.06	50%	193%	0.52	33.5x	0.26	133.8x	0.05	3345x	0.03	13382x
R 1	3	$ 0.143	5.8%	$0.09	50%	84%	1.19	6.3x	0.60	25.2x	0.12	631x	0.06	2524x
R 1	1–6	$ 0.912	1.4%	$0.34	50%	134%	0.75	16.2x	0.37	64.7x	0.07	6939x	0.02	27756x
R 1	1	$ 0.244	4.2%	$0.04	50%	278%	0.36	69.4x	0.37	277.6x	0.07	425x	0.07	1700x
R 1	2–3	$ 0.207	2.3%	$0.12	50%	84%	1.20	6.3x	0.60	25.2x	0.12	629x	0.06	2515x
R 2	1a	$ 0.139	0.9%	$0.09	15%	24%	4.12	0.5x	2.06	2.1x	0.41	53x	0.21	212x
R 2	1b	$ 0.142	0.9%	$0.08	15%	25%	3.99	0.6x	2.00	2.3x	0.40	57x	0.20	226x
R 2	1c	$ 0.131	0.8%	$0.09	15%	23%	4.33	0.5x	2.17	1.9x	0.43	48x	0.22	192x
R 3	1	$ 0.061	9.5%	$0.04	30%	52%	1.92	2.4x	0.96	9.7x	0.19	243x	0.10	972x
R 3	2	$ 0.044	8.0%	$0.04	30%	34%	2.96	1.0x	1.48	4.1x	0.30	103x	0.15	411x
R 3	3	$ 0.065	6.7%	$0.09	30%	22%	4.50	0.4x	2.25	1.8x	0.45	44x	0.23	177x
R 3	4	$ 0.051	7.8%	$0.06	30%	26%	3.82	0.6x	1.91	2.5x	0.38	62x	0.19	247x
R 3	5	$ 0.049	5.5%	$0.07	30%	21%	4.73	0.4x	2.36	1.6x	0.47	40x	0.24	161x
R 3	6	$ 0.064	4.8%	$0.08	30%	25%	3.98	0.6x	1.99	2.3x	0.40	57x	0.20	227x
R 3	7	$ 0.032	10.6%	$0.11	30%	9%	11.32	0.1x	5.66	0.3x	1.13	7x	0.57	28x
R 4	1	$ 0.031	10.9%	$0.13	40%	10%	10.45	0.1x	5.22	0.3x	1.04	8x	0.52	33x
R 5	1	$ 0.215	0.8%	$0.11	30%	57%	1.76	2.9x	0.88	11.6x	0.18	291x	0.09	1165x
R 5	2	$ 0.190	4.4%	$0.39	30%	15%	6.90	0.2x	3.45	0.8x	0.69	19x	0.34	76x

New Accounts Only

		Key Statistical Properties of Campaign					HARD: Ads did anything? H0: ROI=-100% Ha: ROI=0%		HARDER: Worth money? H0: ROI=0% Ha: ROI=50%		HARDEST: To optimize? H0: ROI=0% Ha: ROI=10%		CRAZY: Profits maximized? H0: ROI=0% Ha: ROI=5%	
Adv	#	SE β New Accts	Radius 95% %New Accts	Spend Per Person	Lifetime Value	SE ROI	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3	E[t]	Mult. E[t]=3
F 1	1a	69	15.6%	$0.06	$1,000	138%	0.73	17.1x	0.36	68.3x	0.07	1707x	0.04	6828x
F 1	1b	69	17.7%	$0.07	$1,000	137%	0.73	17.0x	0.36	67.9x	0.07	1697x	0.04	6790x
F 1	1c	70	10.9%	$0.07	$1,000	93%	1.07	7.8x	0.54	31.4x	0.11	785x	0.05	3139x
F 1	1d	70	10.5%	$0.07	$1,000	93%	1.08	7.7x	0.54	30.9x	0.11	774x	0.05	3094x
F 2	1	288	5.5%	$0.03	$1,000	47%	2.13	2.0x	1.06	8.0x	0.21	199x	0.11	795x
F 2	1	46	8.3%	$0.02	$1,000	233%	0.43	48.7x	0.21	195.0x	0.04	4874x	0.02	19496x

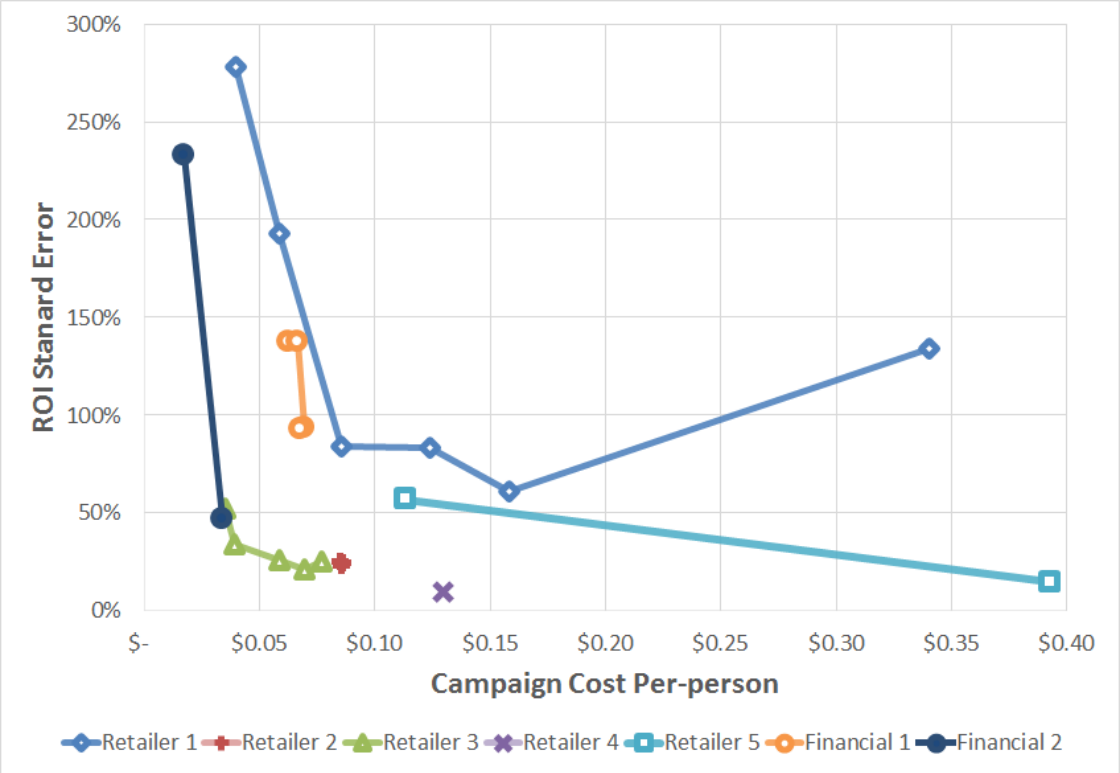


Figure 3: Relationship between ROI uncertainty and campaign cost.

campaign requires a larger impact on sales to deliver the same percentage return. Measured against the same background noise, a larger impact is easier to identify than a smaller one—the more intense experiment allows for better power. This means that identifying the full shape of the $\beta(c)$ function would be very difficult. As one moves closer to the origin, the noise in estimation tends to increase because a cheap campaign results in weak statistical tests.

In the final 8 columns of Table 2, we examine an advertiser's ability to evaluate various sets of hypotheses on the returns to expenditure. We start with disparate null and alternative hypotheses, posing the question "Could the advertiser distinguish 0 effect (-100% ROI) from positive returns (ROI>0%)?" We then draw the hypotheses closer together, to tolerances more typical of investment decisions. For each hypothesis set, we give the expected t-statistic, $E[t]$, to reject the null hypothesis—a natural measure of expected statistical significance when true state of the world is given by the alternative hypothesis. An expected t-statistic of 3 provides power of

91% with a one-sided test size of 5%. We also give a "data multiplier," which tells us how much larger the experiment would have to be in terms of new (independent) individuals, and implicitly the total cost, to achieve $E[t] = 3$ when the alternative hypothesis is true. The experiment could also be made larger by holding N constant and lengthening the duration using the same spend per week. Here we focus on N because it does not require us to model the within person serial correlation of purchases. Naturally if individuals' purchases were independent across weeks, then adding a person-week could be done just as effectively by adding another week to the existing set of targeted individuals.[17]

We start with distinguishing no impact (-100% ROI) from positive returns (ROI> 0%). In fact, most papers on ad effectiveness use this is as the primary hypothesis to test—the goal being to measure whether the causal influence on sales is significantly different from 0 (Bagwell, 2005). The break-even sales impact is inversely related to the firm's margin; for these advertisers the required sales impact to break even is 2–7 times the per-person expenditure (Column 6). Nine of 25 experiments had E[t]<1.65 (Column 9), meaning the most likely outcome was failing to reject -100% ROI when the truth was the ad broke even.[18] Ten of 25 of the experiments had E[t]>3, meaning they possessed sufficient power to reliably determine if the ads had a causal effect on consumer behavior. These tests are performed in the multiple papers cited earlier and generally reveal a statistically significant impact of advertising (Lewis and Reiley, 2010; Johnson et al., 2013; Lewis and Schreiner, 2010). The papers also discuss features of advertising such as the impact of local targeting and the role of age in advertising responsiveness; interested readers are encouraged to consult these papers.

Simply rejecting the null that a campaign was a total waste of money is not a very ambitious goal. In the "harder" column we ask a more appropriate question from a business perspective, "Are the ads profitable?" Here we set the null hypothesis

[17]If the serial correlation is high and positive, then adding more weeks is much less effective than adding more people, if it is high and negative, it is actually more effective. Note also that campaigns are typically rather short because firms like to rotate the advertising copy so that ads do not get stale and ignored.

[18]If E[t]<1.65, even with a one-sided test, more than half the time the t-statistic will be less than the critical value due to the symmetry of the distribution. As an aside, we note that these experiments are not meant to represent optimal experimental design. Often the advertisers came to us looking to understand how much can be learned via experimentation, given a number of budgetary and campaign-objective constraints.

as ROI=0% and the alternative to a blockbuster return of 50%. Here 12 of 25 experiments had E[t]<1 (severely underpowered), 4 had E[t]∈[1,2], 5 had E[t]∈[2,3] (90%>power>50%), and only 3 had E[t]>3. Thus, only 3 of the 25 had sufficient power to reliably conclude that a *wildly profitable* campaign was worth the money, and an additional 5 could reach this mark by increasing the size of the experiment by a factor of about 2.5 (those with E[t]∈[2,3]) or by using other methods to optimize the experimental design.[19] The median campaign would have to be 9 times larger to have sufficient power in this setting.

The most powerful experiments were Retailer 5's second campaign, which cost $180,000 and reached 457,968 people, and Retailer 4's campaign, which cost $90,000 and reached 1,075,828 people. For Retailer 5's second campaign, the relatively high precision is largely due to it having the most intense in terms of per-person spend ($0.39). The precision improvement associated with tripling the spend as compared to an earlier campaign is shown graphically in Figure 3. Retailer 4 had good power due to two key factors: it had the fourth highest per-person spend and the second lowest standard deviation of sales.

Distinguishing a highly successful campaign from one that just broke even is not the optimization standard we typically apply in economics, yet our analysis shows that reliably distinguishing a 50% from 0% ROI is typically not possible with a $100,000 experiment. In the third and fourth columns from the right, we draw the hypotheses closer together to a difference of 10 percentage points. While we use 0% and 10% for instructive purposes, in reality the target ROI would need to be estimated as well. Strikingly, every experiment is *severely* underpowered to reject 0% ROI in favor of 10%. E[t] is less than 0.5 for 21 of 25 campaigns and even the most powerful experiment would have to be 7 times larger to have sufficient power to distinguish this difference. The median retail sales experiment would have to be *61 times larger* to reliably detect the difference between an investment that, using conventional standards, would be considered a strong performer (10% ROI) and one that would be not worth the time and effort (0% ROI). For new account sign-ups at financial service firms, the median multiplier is an almost comical 1241—this reflects the all-or-nothing nature of consumption patterns for these firms, a feature shared by other heavily advertised goods such as automobiles.

[19]These could include a larger control group, longer time period, and so forth.

In the final two columns of Table 2 we push the envelope further, setting the difference between the test hypotheses to 5 percentage points. The expected t-statistics and multipliers for E[t]=3 demonstrate that this is not a question an advertiser could reasonably hope to answer for a specific campaign or in the medium-run across campaigns—in a literal sense, the total U.S. population and the advertiser's annual advertising budget are binding constraints in most cases. These last two hypotheses sets are not straw men. These are the real standards we use in textbooks, teach our undergraduates and MBAs, and employ for many investment decisions. The fact they are nearly impossible to apply for these retailers and financial service providers is a key contribution of the paper, and in the discussion section we use data from industry groups to argue that these advertisers are not atypical. In fact 5% ROI in our setting is over an approximately two week period, corresponding to over a 100% annualized ROI. If were instead to focus on 5% annualized ROI, the problem becomes at least 400 times harder.[20]

Many investment decisions involve underlying certainty. In drug discovery, for example, a handful of drugs like *Lipitor* are big hits, and the vast majority never make it to clinical trials. Drug manufacturers typically hold large diversified portfolios of compounds for this very reason and ex-post profit measurement is relatively straightforward. Advertisers tend to vary ad copy and campaign style to diversify expenditure. This does guard against idiosyncratic risk of a "dud" campaign, but it does not guarantee the firm is at a profitable point on the beta function because ex-post measurement is so difficult. A good analog may be management consulting. Bloom et al. (2013) document the difficulty in measuring the returns to consulting services and conduct the first randomized trial to measure the causal influence of these expensive services. The authors document a positive effect of consulting but also report that precise ROI statements are difficult to make. One might have thought that the ability to randomize over millions of users would set advertising off to an inference advantage over company-wide type choices like management consulting services that are notoriously hard to evaluate, but it turns out these sorts of expenditures are good analogs to advertising.[21]

[20] We are trying to estimate 1/20th of the effect size we previously were, which is 20^2 times harder.

[21] Albeit for different reasons. Consulting should have a large impact (commensurate to the high costs), but it is hard to independently administer within a firm and time trends make it hard to compute before-after counterfactuals.

3.3 Determining the ROI target

Here we briefly touch on how a firm would determine the ROI target in our simple model. Returning to Figure 1, there are 3 important regions, split by c^* and c_h. c^* gives the optimal per-person spend and defines the ROI target: $\frac{\beta(c^*)m-c^*}{c^*}$. c_h gives the break-even point at which average ROI is zero. For $c < c^*$, average ROI is positive but the firm is under-advertising—ROI is too high. For $c > c^*$, the firm is over-advertising, average ROI is still positive as long as $c < c_h$, but marginal returns are negative. We believe ROI estimates in this region would be hardest to act on. Positive ROI means the firm made money on the campaign. It may be psychologically and politically hard to reduce spend when the enterprise was just shown to be profitable. When $c > c_h$ ROI is negative and the actionable insights are much clearer. The firm should advertise less, and politically this would be an easy decision to execute.

One simple way to estimate the sales impact function would be to run an experiment with several treatment in which cost per person is exogenously varied.[22] Each treatment give a estimate in $(c, \beta(c))$ space shown in Figure 1. Our analysis shows that each of these points would have large confidence intervals, so fitting the function with non-parametric techniques that take into account statistical uncertainty would result in a wide range of ROI curves that cannot be rejected, providing little guidance for the advertiser, especially at low levels of per-person spend.

Instead, a firm may use simple comparisons to measure marginal profit. Consider two spend levels $0 < c_1 < c_2$. Marginal profit is given by $(\beta(c_2)m - c_2) - (\beta(c_1)m - c_1) = (\beta(c_2) - \beta(c_1))m - (c_2 - c_1)$. Just as with measuring the average ROI where $c_1 = 0$, this boils down to comparing two estimates with large confidence intervals. However, the marginal ROI situation is much worse because the cost differential between the two campaigns, $\Delta c = c_2 - c_1$, is smaller when $c_1 > 0$. This has two implications. First, the expected effect on profits will be much smaller due to the decreasing returns of c in $\beta(c)$. Second, the variance of marginal ROI has the cost differential in the denominator:

$$Var(ROI(\Delta c)) = \left(\frac{m}{\Delta c}\right)^2 Var\left(\beta(c_2) - \beta(c_1)\right). \tag{9}$$

[22] See Johnson, Lewis, and Reiley (2013) for an example.

If Δc is small as we try to optimize ROI in the neighborhood of c^*, the variance of our ROI estimates diverge to infinity. By definition, $\Delta c \approx 0$ for marginal campaigns, but all campaigns are marginal by construction as they represent an elemental unit of broader marketing strategies. Smaller cost differences push our effective cost per user down, meaning the points on the left side of Figure 3 are more representative of the hypothesis tests we will run with small Δc. Ideally, we would like to get to a point where this marginal profit estimate is zero (or equal to the cost of capital), but achieving such precise estimates is essentially impossible.

Pulling out a simple first derivative is much more difficult than estimating a single point. In the real world various other factors further complicate matters. Concerned with ad copy "wear out," firms tend to use different/new ad copy for different campaigns (Eastlack Jr and Rao, 1989; Lewis, 2010). Comparing campaigns of differing intensity and ad copy adds a non-trivial economic wrinkle. Using the same logic as above, determining if two creatives are significantly different will only be possible when their performance differs by a very wide margin. If creatives show considerable differences in user impact, then aggregating data across campaigns is naturally less useful—we are back to evaluating campaigns in isolation, which we have just shown is a very difficult enterprise.

4 Discussion

In this section we will first address the generalizability of our findings. We then discuss the implications on the marketplace.

4.1 Generalizing our findings: Representativeness of the sample

A natural concern is that our selected sample of advertisers, namely those who agreed to data sharing and experimentation with a large web publisher, are not representative of the advertising market as a whole. We address this concern here by stress testing our results from a number of different angles concerning the representativeness of both the sample and advertising medium.

4.1.1 Do these firms have unusually high sales volatility?

Sales volatility and the required response rate are the key factors governing the difficulty in measuring the returns to advertising. The required response rate to achieve a target ROI scales linearly with cost-per-user. The ads in our study were representative of premium online display, about 1/3 the price per impression of a 30-second TV commercial, but with more impressions delivered per user—these campaigns have a cost-per-user that is common in the industry. To get an idea of how the sales volatility of our firms compares to other heavily advertised categories, we use data from an industry that advertises heavily and for which data are available: American automakers.[23]

We back out sales volatility using published data and a few back-of-the-envelope assumptions. It is reasonable to suppose the average American purchases a new car every 5–10 years. We will generously assume it is every 5 years (a higher purchase frequency makes inference easier). Suppose that the advertiser has a 15% market share. Then the annual probability of purchase for this automaker is 0.03 (Pr(buy) = .2*.15 = .03), which implies a standard deviation of $\sqrt{0.03} \approx \frac{1}{6}$. To turn this into a dollar figure, we use the national average sales price for new cars, \$29,793.[24] Mean annual sales per-person is \$893 (μ=\$29,793$*0.2*0.15$, price $\times$ annual purchase rate $\times$ market share) and σ= 1/6*\$29,793 = \$4,700. This gives a $\frac{\sigma}{\mu}$ ratio of roughly 5, similar to our finding of 10 for retailers. However, this is *yearly*, as opposed to the finer granularity used in our study. To convert this into a monthly figure, we multiply by $(1/\sqrt{12})/(1/12) = \sqrt{12}$, yielding a ratio of 20:1.[25]

[23]The leading industry trade group, the National Automobile Dealers Association (NADA), estimates that automakers spend \$674 in advertising for each vehicle sold.

[24]Source: http://www.nada.org/Publications/NADADATA/2011/default.

[25]Here we assumed zero variance in purchase price. In this setting, including variation in price does not make the inference problem much more difficult—most of the difficulty is driven by rarity of purchases. More generally, individual sales equals probability of purchase times the dollar value of the purchase. The variance of each component contributes to overall sales volatility as given by: Let $Y = p * \$$ where p is purchase probability and \$ is basket size. $\frac{\sqrt{var(Y)}}{E[Y]} = \frac{\sqrt{E[p]^2 Var(\$) + Var(p) E[\$]^2 + Var(\$) Var(p)}}{E[p]E[\$]}$. Both components can presumably be impacted by advertising (getting customers to come to the store and getting more items the in the shopping cart). For the retailers in our study, back-of-the-envelope calculations indicate that each component accounts contributes significantly to the total, but purchase rarity probably accounts a larger portion than variation in basket size conditional on purchasing. For example, using values from Lewis and Reiley (2008), calculations show that ignoring the components with $Var(\$)$ reduces the total coefficient of variation by about 40%. Ignoring the basket size component would make the problem somewhat easier statistically, but would induce a bias of unknown size.

Heavily advertised categories such as high-end durable goods, subscription services such as credit cards, and infrequent big-ticket purchases like vacations all seem to have consumption patterns that are more volatile than the retailers we studied selling sweaters and dress shirts and about as volatile as the financial service firms who also face an "all-or-nothing" consumption profile. Political advertising appears to share similar difficulties (Broockman and Green, ming).[26]

4.1.2 Are these campaigns too small?

Our scale multipliers give an idea of the cost necessary to push confidence intervals to informative widths—the implied cost (if that many unique individuals were available) was often in the tens of millions of dollars, far more expensive than even the largest reach advertisement in the US, the NFL Super Bowl, which we will use here in a thought experiment, supposing that the 30-second TV spots can be individually randomized. We will try to define the set of advertisers that can both afford a Super Bowl spot and detect the return on investment.

The affordability constraint is simply an accounting exercise to ensure firm's advertising budget can accommodate such a large expenditure. To build intuition on the detectability constraint, recall that ROI is the percentage return on the *ad cost*—it does not depend on the baseline level of sales. The sales *level* lift that nets a positive ROI is a much larger *percentage* lift for a small firm than for larger firms and thus more likely to stand out statistically. The "detectability constraint" gives the largest firm, in terms of annual revenue, that can meaningfully evaluate a given ROI hypothesis set.

In consideration of space, we put the formal argument in the Appendix. We set the analysis window $w = 2$ (weeks) to match most of the analysis of this paper. $t_{ROI} = 3$ to match our standard power requirement and $\frac{\sigma}{\mu} = 10$ to match the value we see strong evidence for in our study, even though it will understate volatility for advertisers such as automakers and financial service firms.[27] For the advertising

[26]They run a small experiment with 1,400 surveyed Facebook users, which places the lower bound of the marginal cost of a vote at \$0.50 and the upper bound at infinity (since they cannot reject zero effect of the ads). To evaluate a reasonable vote cost figure of around \$50, a value we take from spending in swing states in presidential elections, their experiment would need to be scaled by a factor of 30,000, to 400,000 unique users, to reliably reject a cost of \$50 if the ads, in fact, have no effect. Even this coarse test would not be feasible for many candidates in many elections.

[27]We use $\rho = 0.5$ to match the empirical viewing share for adults for the Super Bowl. See

budget we choose a value, 5% of revenue, which exceeds advertising budgets for most major firms.[28] We report bounds for two values of gross margin: 0.25 and 0.50. The final step is to calibrate pricing and audience. We use the following parameters: N_E is 50 million (1/2 the viewers) and the cost of the ad is 1/2 the market rate, $C = \$1,000,000$ (1/2 the cost).

Table 3 gives the upper and lower bounds on annual revenue (what really matters is revenue for the two week evaluation period, but we opted to report the more interpretable annual figure, implicitly assuming sales are relatively smooth across time). If an ad promotes only a specific product group, for instance the 2011 Honda Civic, then the relevant figure to compare to the bounds would be the revenue for that product group. Examining Row 1, we see that most companies would be able to reliably determine if the ad causally impacted consumers. Major automobile manufacturers (which are low margin) doing brand advertising would exceed this limit, but specific model-years fall below it.[29]

Table 3: Super Bowl "Impossibility" Theorem Bounds

H_A: ROI	H_0: ROI	Affordability Annual Rev.	Detectability, m=.50 Annual Rev.	Detectability, m=.25 Annual Rev.
0%	-100%	\$2.08B	\$34.47B	\$63.3B
50%	0%	\$2.08B	\$17.33B	\$34.6B
10%	0%	\$2.08B	\$3.47B	\$6.9B
5%	0%	\$2.08B	\$1.73B	\$3.4B

We see in Row 2 that many companies and product categories could reliably distinguish 50% ROI from 0%—the bounds are \$17.3 billion and \$34.6 billion for the high and low margins respectively—but large firms or products could not. For the final two hypothesis sets, the bands are tight to vanishing. It is nearly impossible to be large enough to afford the ad, but small enough to reliably detect meaningful differences in ROI.

Appendix for more details.

[28] Source: Kantar AdSpender.

[29] However, we have assumed a $\frac{\sigma}{\mu}$ ratio of 10, which is probably half the true value for car sales over 2-4 week time frame, meaning the correct bound is probably twice as high.

4.1.3 Are these campaigns not targeted enough?

A targeted ad is designed to reach a customer with preferences towards the product, and empirical evidence supports the view that targeted ads have a stronger influence on purchasing behavior (Montgomery, 1997; Rossi et al., 1996). Can a firm more powerfully assess their advertising stock by performing experiments on the particularly susceptible portion of the population? The trade-off is that targeting reduces the size of the experiment, which works against power at a rate of $\sqrt{N}$, but increases the expected impact, making ROI easier to detect.

Suppose there are N individuals in the population the firm would consider advertising to. We assume that the firm does not know how a campaign will impact each individual, but can order them in expectation. The firm wants to design an experiment using the first M of the possible N individuals. We define $\Delta\mu(M)$, $\sigma(M)$, and $c(M)$ as the mean sales impact, standard deviation of sales, and average cost functions, respectively, when advertising to the first M people. The t-statistic against the null hypothesis of -100% ROI is given by: $t = \sqrt{\frac{M}{2}} \cdot \frac{\Delta\mu(M)}{\sigma(M)}$.

Assuming constant variance and taking the derivative with respect to M we get:

$$\frac{dt}{dM} = \frac{\Delta\mu(M)}{\sigma(M)} \frac{1}{2\sqrt{2M}} + \sqrt{\frac{M}{2}} \frac{\Delta\mu'(M)}{\sigma(M)} \tag{10}$$

If the ad has a constant effect on the population, then $\Delta\mu'(M) = 0$ and we get the standard results that t increases at a rate proportional to $\frac{1}{\sqrt{M}}$. With targeting, $\Delta\mu'(M) < 0$. Simplifying the right hand side, we find the t-statistic is increasing in M if the targeting effect decays slower than $\frac{\Delta\mu(M)}{2\sqrt{2M}}$. Thus, the question of whether targeting helps or hurts inference is an empirical one. If the sales impact is concentrated on a certain portion of the population, one is better off reducing sample size to gain a higher signal-to-noise ratio. Conversely, if influence is spread rather evenly across the population, targeting damages power. Additional details of this argument are in the Appendix.

These same results apply to search advertising. N is the total population, M is the number of customers who see search ads, and $c(M)$ is the average cost per search ad impression.[30] By restricting attention to the sales of those M users who saw a

[30] $c = CPC \cdot CTR$ per search ad impression per user where CPC is the cost per search ad click and CTR is the click-through rate on the search ad.

search ad, we increase[31] t (CITE?: Lewis 20?? SMALLEST SUPERSET or GHOST ADS papers). However, current search advertising technology does not admit linking individual-level search ad exposure to individual-level sales. Thus, we should expect search advertising effectiveness measurement to exhibit two settings: 1) advertisers who primarily receive traffic through search advertising can readily measure the effectiveness search ads because virtually all sales are due to search advertising, and 2) all other search advertisers could measure search advertising effectiveness better but cannot because they are unable to restrict attention to customers who were exposed to search advertising. This fact, combined with each advertiser's unique situation, helps understand why we should expect wide heterogeneity in advertisers' uncertainty regarding their search advertising spending.

4.1.4 Would longer measurement windows help?

Any analysis of the returns to advertising invariably has to specify the window of time to be included in the study. We followed the standard practice of the campaign period and a relatively short window after the campaign ended. Perhaps by adding more data on the time dimension, we would get a better estimate of the cumulative effect and statistical precision would improve. It turns out that if the effects of advertising decay over time, a natural assumption, then adding additional data will at some point only cloud inference.

We present the formal argument in the Appendix. The key proposition is the following:

> *If the next week's expected effect is less than one-half the average effect over all previous weeks, then adding it in will only reduce precision.*

The proposition tells us when a marginal week hurts estimation precision because it introduces more noise than signal. As an example, suppose the causal impact of the advertising on weeks 1, 2, and 3 is 5%, 2%, and z, respectively. Then z must be greater than $\frac{5+2}{2} = 1.75$. In other words, unless there is very limited decay in the ad effect over time, we would be better off curtailing the evaluation window to two weeks. With moderate decay, optimal evaluation windows (from a power perspective) get quite short. An additional week of data increases the effective

[31] This should read "weakly increase" if there are no sales by unexposed users.

sample size and the cumulative impact, but reduces the average per-time-period impact, watering down the effect we are trying to measure.

The proposition can provide helpful guidance and helps explain why short windows are generally used, but quantitatively applying it requires precise ROI estimates for the very inference problem we are trying to solve. So, in the end, the practitioner and econometrician alike must make a judgment call[32], and right now our judgment is to use 1-4 week windows. This is an unsatisfying step in the estimation process for any empirical scientist, but it is necessary because estimating the long-run effect of an advertising campaign is a losing proposition.

4.1.5 Supportive evidence from an advertiser's research

We think it is natural for an economist to be skeptical of our claim that advertisers largely do not (and often *cannot*) know the effectiveness of their advertising expenditures. Recent work by eBay Research Labs (Blake, Nosko, and Tadelis, 2013) offers an illuminating perspective from the advertiser's side on this claim.

eBay has historically been one of the largest buyers of paid links on search engine results pages. Experimentation can be done relatively easily for paid search listings by "pausing" the ad at pre-specified times, yet prior to this research the company apparently had not run any such experiments. Given the hundreds of million of dollars it had spent on such listings, the lack of experimentation might come across as a bit of a shock, but it is far less surprising viewed in the light of a market in which participants do not expect much feedback.

The authors examine the returns to branded keywords (e.g., "tablet computer ebay") and unbranded keywords (e.g., "tablet computer"). For branded terms, they use pause experiments to show that most of the clicks on paid search links would have otherwise occurred on an "organic" link, which was typically right below the ad in the algorithmic results. For unbranded terms, the authors turn search ads on and off for geographic regions. They estimate that paid search is causally linked to 0.44% of total sales, with a standard error of 0.62%, leading to a 95% confidence interval of (-0.77%, 1.66%). The 0.44% sales impact corresponds to -68% ROI, a considerable loss; however, the top of the confidence interval is +16% ROI, meaning

[32]It turns out that these types of judgment calls look a lot like estimating an endogenous structural break, whose estimated location is super-consistent (Bai and Perron, 1998). So, these judgment calls are not quite as statistically tenuous as they might first appear.

they could not reject profitability at standard confidence levels.

The authors examine the impact of varying ad spending across geographical regions by regressing sales revenue on search spending using the randomization as the instrument. Ordinary Least Squares, even with a full set of controls, grossly overstates the true impact due to temporally varying purchase intent, a bias first documented in Lewis et al. (2011). The Instrumental Variables estimate is that a 10% increase in spending is associated with a 0.56% increase in sales. But the 95% confidence interval is about 30 times wider than the point estimate: (-8.1%, 8.7%).

We think this paper dovetails with our results nicely. First, the authors are employed by a large advertiser and openly claim the company did not know the returns to advertising and strongly imply (p. 14, paragraph 2) that observational methods were being used that severely overstated returns. Second, the experiment confirms that truly ineffective campaigns can be identified via large scale experimentation (but not observational methods). Third, for the more nuanced case of unbranded keywords, the estimates on the marginal dollar spent have enormous confidence intervals, and the considerably smaller confidence interval on the total ROI is over 100 percentage points wide.

4.2 Generalizing our findings: Improving experimental design and data collection

Two additional potential criticisms of our findings are that the experiments we evaluated may not have possessed state-of-the-art design and that data collection methodology and infrastructure will quickly improve to overcome the difficulties we have identified. We address these concerns in this subsection, noting where we think improvements will be made and where fundamental inference difficulties will likely remain..

4.2.1 Optimizing ex-post evaluation and experimental design

In Section 2.2 we calibrated an advertiser's inference problem using univariate linear regression for expository clarity. In our actual estimation, we conditioned on the user level covariates listed in the column labeled by the vector **W** in Table 1 using several methods to strengthen power, such panel techniques to predict and absorb residual

variation. We find that lagged sales is the best predictor and used it where possible– it's strongest when predicting long campaigns with long histories, reducing variance in the dependent variable during the test period by as much as 40%. Demographics are typically not found to meaningfully improve power. A recent paper in the computer science literature on improving the power of online experiments, Deng et al. (2013), also finds that including lagged dependent variables (pre-determined before the start of the experiment) can reduce residual variation by as much as 50%. From a design perspective, a related method to including pre-determined (lagged) dependent variables in the analysis is to match like users into pairs and then randomize exposure within each pair. This sort of pre-experiment matching is more useful for small samples as it can unsure against an unlucky randomization step—with large samples this is highly unlikely to occur.

We have alluded to another technique to improve power throughout: concentrated tests. Figure 3 shows how within a given firm, larger expenditures per-person (conditional on experiment size) are associated with lower standard errors on the ROI estimate. Intuitively a larger spend creates more variation in the X variable for which we are trying to identify a causal impact. In our discussion of targeting in the last section, we show that a firm can sometimes improve power by using the most susceptible part of the population in the experiment. Similarly, if a portion of the customer base had lower variance in sales over time, these customers are attractive to include in an experiment. In these ways, firms can carve out a portion of the population with statistically favorable properties. The problem of course the measurements of the returns to advertising for these sub-samples may not reflect returns for the population of consumers the firm wishes to advertise too. Moreover, a concentrated test (high per-person spend) may be at a level of advertising beyond the firm's optimum. Our conclusion is thus that while these methods may be offer improvements in power for some firms in some situations, they certainly do not the fundamental inference difficulties we have raised in this paper.

The final design advance on the horizon is free control ads for experiments, sometimes referred to as "ghost ads." In the past, advertisers would have to pay for both treatment and control ads (often an ad for a charity), so that they could measure who actually saw a control ad (served with the same targeting characteristics, etc., as a treatment ad), forming the comparison group. Technology now exists to hold the control group out from the advertiser's ad but measure when an attempt to

serve the treatment ad was made. In each of these cases, the treatment ad would have been shown had the user now been in the control group. The system records what ad was shown, so one can control for possible competitive effects of advertising, but this does add a non-trivial wrinkle to the analysis. Even so, the prospect of free control ads means doubling experiment size at the same cost is feasible. Examining the multipliers in Table 2, we see this will help this class of advertisers running $100,000 experiments infer if ad expenditure causally impacts consumers. An improvement of 2x will in general not be much help evaluating any of the more realistic hypotheses sets, but we do note that any reduction in experimental costs will only promote accurate inference in the marketplace.

4.2.2 Advancing scale

a. Bigger publishers and experimentation platforms b. Linking data across experiments with unique identifiers

4.2.3 Aggregating across experiments and forming priors

a. Problem: if we want to evaluate if a campaign was worth it, this punts on that question entirely b. Gets at the global average question c. Form a useable prior to evaluate future experiments

4.3 Marketplace implications

4.3.1 Average vs. marginal ROI

In textbooks, the distinction between average and marginal is unambiguous. "Average" is just the total sales increase divided by total spend and "marginal" is the impact of that "last little bit" of advertising, divided by its cost. In addition to advertising online, most of our firms were actively advertising on television, out-of-home, and through direct mailings. Exactly what part of this spend is marginal, from the perspective of the firm's decisions, is entirely unclear. Mechanically, any experiment in a single channel is a evaluating a "marginal campaign," because the control subjects still see the same billboards, television commercials, etc.

Suppose a firm runs a series of experiments in given medium and eventually rejects 0% ROI in favor of the most likely alternative, say -50% ROI. What is

the appropriate response? Spending should be cut, but where? The online ads could have lacked the impact to break even. Or maybe the firm is advertising too heavily across the board, and all spending should be cut equally. Marginal experiments could signal the effectiveness of spend generally, provide idiosyncratic feedback on a particular campaign, a particular cross-media combination or any combination therein. The marginal-average interpretation adds a whole new layer of complexity—techniques will likely be developed to address this complexity as cross-media experimentation becomes more feasible.

4.3.2 A new competitive advantage of scale

An implication of the low power of advertising experiments is that large publishers have an advantage not only through the classical notion of having larger reach, but also by having the user base to run reliable experiments. Table 2 shows that some large advertisers could narrow confidence intervals to an acceptable tolerance with experiments in the tens of millions of users in each treatment cell. Only the largest publishers could offer such a product. If experimentation becomes more common, a trend we believe is occurring, scale will confer a new competitive advantage. A respectable but smaller publisher simply cannot provide feedback of the same quality as a massive publisher and may be better off outsourcing ad-serving to a larger network. For smaller advertisers, the large publisher can leverage its scale to recommend ad features based on findings from past experimentation with larger firms. Increased experimentation thus has the potential to fundamentally shape the organization of web publishing and potentially other advertising-based industries.

4.3.3 Reputation, moral hazard and communication within the firm

Table 2 shows that evaluating an individual campaign is difficult to do with any reasonable level of precision. Feedback on the returns to advertising accumulates slowly. Given that a firm cannot initially verify quality of services provided, reputation among advertising agencies becomes more important and can represent a barrier to entry from upstarts. This will only occur if performance is verifiable in the long-run (so that good reputation means something), but cannot be fully contracted on (otherwise the upstart could write a contract to insure against poor performance).

A related issue is that incentives can create a moral hazard problem for reporting ROI estimates truthfully. We call the person responsible for purchasing a specific campaign the "media buyer." The media buyer reports to her manager, who is in a chain of intermediaries that eventually reports to entity that cares about the truth. The media buyer gets a bonus based the manager's posterior belief on campaign ROI. If reports are delivered with certainty and completely verifiable, there is no agency problem. If they are totally unverifiable, we are in a *cheap talk game* (Crawford and Sobel, 1982) where strategic communication leads to reports that are correlated with the agent's signal (the estimate), but noisy due to the common knowledge of the agent's bias.[33] Since it is very hard to disprove a report with other data and estimates themselves are noisy and likely manipulatable[34], we contend there is room for selective filtering as in cheap talk equilibria.[35] A recent paper addressed cheap talk in hierarchical networks and showed that the informativeness of pure strategy equilibrium communication is limited by the most biased agent in the reporting chain, who forms a choke point (Ambrus et al., 2013).

Alternatively we might view the strategic communication as *persuasion game* (Milgrom and Roberts, 1986) in which the media buyer wants to persuade her manager that her annual ad spending delivered a high ROI. Applying the model of Shin (1994), we suppose that the manager is unsure which campaigns have verifiable ROI estimates. In equilibrium, the manager will be skeptical because the media buyer will report good news when available but filter bad news, which limits the amount of information that can travel of up the chain of command.

4.3.4 Variance in advertising spend across competitors

Information is scarce in the advertising market, meaning that the "selective pressure" on advertising spending is weak. We would thus expect significant heterogeneity in advertising spend by similar firms in the same industry. Empirically testing this prediction is difficult because there are many economic reasons firms could have

[33]There are always "babbling" equilibria, which seem rather unreasonable in this setting.

[34]This can be done by varying the technique used, changing the control variable set to find the highest point estimate, etc. With a fragile estimate, these techniques can be effective.

[35]If the principle could access the raw data at some cost it could mitigate this problem. As a practical matter, it is unclear who would do this. The proximate manager probably wants to have a positive report almost as much as the agent. The chief marketing officer probably wants the truth, but cannot possibly verify all reports.

different advertising strategies.[36] We thus limit our comparisons to industries dominated by a handful of firms that share key characteristics reported to the SEC such as margins, access to technology, annual revenue, and customer base. Our data on advertising expenditure comes from Kantar Media's AdSpender report. Advertising expenditure, revenue, and margins are only available at the firm level, so large conglomerates cannot be used.[37]

In Appendix Table 1 we give advertising expenditure, revenue, and margin for the following U.S. industries: rental cars, mobile phone carriers, international airlines, online financial services, and fast food. These constitute the markets that met the requirements we have laid out and had data availability.

The data show distinct high/low advertising strategies. Advertising expenditure as a percent of revenue differs by a factor of 5 or more between similar firms in the same industry. However, we do not observe this in every industry. For mobile phone carriers, advertising expenditure as a percentage of revenue varies by only a factor of 2: 1.36% (AT&T) to 2.75% (T-Mobile). T-Mobile is 85% smaller, revenue-wise, than AT&T, so it is hard to make much of the two-fold difference—it could be easily explained by concavity in the impact function. A similar pattern emerges with automakers, where advertising expenditure differs by roughly a factor of 2 for Toyota, Honda, Ford, and Hyundai. So while we do see significant dispersion in ad spending in these markets, it is arguably explainable by economic factors.

The remaining industries all show much greater dispersion. Airlines report similar margins and the firms are of similar size, yet advertising differs by a factor of 3 per dollar of revenue. Delta and United are nearly exactly the same size, yet Delta spends about twice as much on advertising, despite reporting a *lower* margin. Rental car companies have nearly identical margins. Three of the 4 major firms advertise between 0.43–0.61% of revenue. The outlier is Dollar Thrifty, which pursues a low-advertising strategy, spending only 0.01% of revenue on advertising—there are no observable characteristics that can explain this difference. Online brokerages

[36]For example, low-cost retailers might compete primarily on price and advertise very little because it erodes slim margins. As we saw in section 2, the lower a firm's margin, the higher the impact of ad has to be to break even.

[37]For instance insurance companies advertise heavily, offer similar services, and have similar margins but are typically part of large conglomerates in order to diversify risk. Kantar's figures should be viewed as informed estimates. Although advertising is tax-deductible and thus reported precisely on corporate tax returns, the IRS does not make this data available to tax-payers. Contact your congressperson...

ScottTrade, TD Ameritrade, and ETrade have similar business models and report identical gross margins. ETrade pursues a high advertising strategy, with 12.63% of revenue going to advertising. ScottTrade spends 8.45% and TD Ameritrade pursues a low-advertising strategy, *6.93 times less* than ETrade per dollar of revenue. In fast food, based on revenue and margins, Wendy's, Burger King, Dairy Queen, and Jack-in-the-Box form a fair comparison set. Burger King and Wendy's use a high advertising strategy—about 12% of revenue, while Dairy Queen and Jack-in-the-Box use a low advertising strategy—about 3% of revenue.

The evidence we present in this subsection is by no means conclusive. However, we think the existence of vastly different advertising strategies, expenditures varying by a factor of 5 or more, by seemingly similar firms operating in the same market with similar margins is consistent with our prediction that vastly different beliefs on the efficacy of advertising are allowed to persist in the market.

4.3.5 How unusual is this market?

In markets with limited informational feedback as to the efficacy of the product, sellers may have a customer base that holds fundamentally incorrect beliefs. Here we look at a two industries that we think share this feature.

The first is management consulting. Bloom et al. (2013) argue that consulting expenditures are rarely implemented in a way that the relevant counterfactual can be formed. To overcome this endogeneity problem, the authors ran a controlled experiment and documented a positive impact of consulting services, but also documented that making precise ROI statements is incredibly difficult.

The second is the vitamin and supplement market. The industry grosses about $20 billion annually, yet it is a contentious point in the medical community as to whether supplements do *anything* for a healthy individual (the main customer base).[38] The effect is supposed to be subtle, making it difficult to reliably detect with any single individual, and across people, medical outcomes that are easily and accurately quantifiable, such as illness requiring hospitalization, are noisy. Observational

[38]Supplements are supposed to improve health for a *healthy person*, not prevent vitamin deficiency diseases such as rickets and scurvy, because in the developed world one gets enough of these vitamins through even the unhealthiest of diets. In a survey of Canadian pediatricians, the overall annual incidence rate for rickets was 2.9 cases per 100,000 people (Ward et al., 2007). In comparison, cancer incidence in Canada is 410.5 cases per 100,000 people (Marrett et al., 2008).

methods suffer from a similar selection bias created by targeting in advertising—people who take supplements may be more health conscious than average or may have recently experienced poor health. Experiments are the natural solution. The Physicians Health Study II (Lee et al., 2005) followed 39,876 healthy women over 12 years. Half received vitamin E through a supplement; the other half took a placebo. The 95% confidence interval on the impact on heart attacks ranged from a 23% risk reduction to an 18% risk increase. We can translate this uncertainty into an "economic confidence interval" using a recent estimate placing cost of a heart attack around $1 million (Shaw et al., 2006). The economic confidence interval is $192 million wide—a whopping *100 times* the $2.1 million cost of vitamins for the study. The economic confidence interval for cancer was of a similar magnitude.[39] The largest experiment to date was thus thoroughly uninformative to dissuade believers in vitamins efficacy or convince non-believers.

5 Conclusion

In this paper we quantitatively assessed the difficulty of the statistical problem facing an advertiser. The challenge is driven by two key facts. First, since campaigns typically involve a modest spend per person, the implied break-even per capita effect of an advertising campaign is small. Second, on the individual level, the ratio of the standard deviation of sales to the mean is about 10:1 for the majority of advertisers we study across a variety of industries.

Using data from 25 large field experiments, accounting for $2.8 million in advertising expenditure, we show that even large experiments can be underpowered, given the noise in sales. A well designed experiment can be informative, but even if true effect of the campaign is economically successful, such as ROI=50%, we show it is difficult, but not impossible, to reliably reject that the campaign merely broke even. More precise tests, such as +/-10% ROI, are shown to be nearly impossible to reliably evaluate. Given the underwhelming power of experiments, the temptation is to turn to observational methods. It turns out that the data features

[39]Cancer incidence ranged from a 7% risk reduction to an 8% increase. A related large-scale experiment, The Selenium and Vitamin E Cancer Prevention Trial, followed 35,533 men from 2001–2008, using a similar design. The results, reported in Lippman et al. (2009), are very similar to Lee et al. (2005).

that make experiments underpowered, severely bias observational methods in this setting. Sources of endogeneity, unaccounted for by observational methods, that explain only a tiny fraction of variation in sales (i.e., R^2 on the order of 0.000005), severely bias estimates, typically upwards due to the nature of ad targeting.

We believe these findings have deep industrial organization implications. First, the advertising market as a whole may have incorrect beliefs about the causal impact of advertising on consumer behavior. As experimentation becomes more common and some firms commit the resources to run the massive (or many large, repeated) experiments necessary to generate informative signals, there could be a meaningful shift in advertising prices. Second, we documented that advertising spending can vary widely across similar firms in the same industry—consistent with the story that signals are weak and priors tend to dominate decision making. Third, the requirement for huge sample sizes in experimentation sets the largest publishers off to an advantage—if the market begins to demand information, their scale will pay an "informational dividend." Overall, the data landscape in the advertising market means decision-making differs fundamentally from our standards notion of profit maximization.

References

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

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

Ambrus, A., Azevedo, E. M., and Kamada, Y. (2013). Hierarchical cheap talk. *Theoretical Economics*, 8(1):233–261.

Bagwell, K. (2005). The economic analysis of advertising. *Handbook of Industrial Organization Volume 3*.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78.

Blake, T., Nosko, C., and Tadelis, S. (2013). Consumer heterogeneity and paid search effectivness: A large scale field experiment. *NBER Working Paper*,

pages 1–26.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? Evidence from India. *The Quarterly Journal of Economics*, 128(1):1–51.

Broockman, D. E. and Green, D. P. (Forthcoming). Do online advertisements increase political candidates' name recognition or favorability? Evidence from randomized field experiments. *Political Behavior*.

Card, D. (1999). The causal effect of education on earnings. *Handbook of Labor Economics*, 3:1801–1863.

Carroll, V., Rao, A., Lee, H., Shapiro, A., and Bayus, B. (1985). The Navy enlistment marketing experiment. *Marketing Science*, 4(4):352–374.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.

Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132. ACM.

Eastlack Jr, J. and Rao, A. (1989). Advertising experiments at the Campbell Soup Company. *Marketing Science*, pages 57–71.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *The American Economic Review*, 97(1):242–259.

Fulgoni, G. and Morn, M. (2008). How online advertising works: Whither the click. *Comscore.com Whitepaper*.

Johnson, G., Lewis, R., and Reiley, D. (2013). Add more ads? experimentally measuring incremental purchases due to increased frequency of online display advertising. *Working paper*.

Kaiser, H. (2005). *Economics of Commodity Promotion Programs: Lessons from California*. Peter Lang Publishing.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620.

Lee, I.-M., Cook, N. R., Gaziano, J. M., Gordon, D., Ridker, P. M., Manson, J. E., Hennekens, C. H., and Buring, J. E. (2005). Vitamin E in the primary prevention of cardiovascular disease and cancer. *The Journal of the American Medical Association*, 294(1):56.

Lewis, R. A. (2010). *Where's the "Wear-Out?": Online Display Ads and the Impact of Frequency*. PhD thesis, MIT PhD Dissertation.

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

Lewis, R. A., Rao, J. M., and Reiley, D. H. (Forthcoming). chapter Measuring the Effects of Advertising: The Digital Frontier. NBER Press.

Lewis, R. A. and Reiley, D. H. (2010). Does retail advertising work: Measuring the effects of advertising on sales via a controlled experiment on Yahoo!

Lewis, R. A. and Schreiner, T. A. (2010). *Can Online Display Advertising Attract New Customers?* PhD thesis, MIT Dept of Economics.

Lippman, S. M., Klein, E. A., Goodman, P. J., Lucia, M. S., Thompson, I. M., Ford, L. G., Parnes, H. L., Minasian, L. M., Gaziano, J. M., Hartline, J. A., et al. (2009). Effect of selenium and vitamin E on risk of prostate cancer and other cancers. *The Journal of the American Medical Association*, 301(1):39–51.

Lodish, L., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., and Stevens, M. (1995). How TV advertising works: A meta-analysis of 389 real world split cable TV advertising experiments. *Journal of Marketing Research*, 32(2):125–139.

Lovell, M. (2008). A simple proof of the FWL theorem. *The Journal of Economic Education*, 39(1):88–91.

Marrett, L., De, P., Airia, P., and Dryer, D. (2008). Cancer in Canada in 2008. *Canadian Medical Association Journal*, 179(11):1163.

Milgrom, P. and Roberts, J. (1986). Relying on the information of interested parties. *The RAND Journal of Economics*, pages 18–32.

Montgomery, A. (1997). Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Science*, pages 315–337.

Rossi, P., McCulloch, R., and Allenby, G. (1996). The value of purchase history data in target marketing. *Marketing Science*, pages 321–340.

Shaw, L. J., Merz, C. N. B., Pepine, C. J., Reis, S. E., Bittner, V., Kip, K. E., Kelsey, S. F., Olson, M., Johnson, B. D., Mankad, S., et al. (2006). The WISE study: The economic burden of angina in women with wuspected ischemic heart disease results from the NIH National Heart, Lung, and Blood Institute sponsored womens ischemia syndrome evaluation. *Circulation*, 114(9):894–904.

Shin, H. S. (1994). News management and the value of firms. *The RAND Journal of Economics*, pages 58–71.

Ward, L. M., Gaboury, I., Ladhani, M., and Zlotkin, S. (2007). Vitamin D-deficiency rickets among children in Canada. *CMAJ*, 177(2):161–166.

Wilbur, K. (2008). How the digital video recorder (DVR) changes traditional television advertising. *Journal of Advertising*, 37(1):143–149.

6 Appendix

6.1 Super Bowl Impossibility Theorem

We will now present the formal argument and calibrate it with data from our experiments and publicly available information on Super Bowl advertising. We need to define some terms. Let N_{Total} be the total adult population, N be the total adult audience, and $\rho = \frac{N}{N_{Total}}$ be the reach of the Super Bowl. N_E gives the number of reached (exposed) individuals; we set $N_E = N/2$ to maximize power. On the cost side, C is the total cost of the ad, and c is the cost per exposed person. Let μ equal the mean purchase amount for all customers during the campaign window and σ be the standard deviation of purchases for customers during the campaign window. We will use $\frac{\sigma}{\mu}$, the coefficient of variation, which we have noted is typically 10 for advertisers in our sample and greater than 10 in other industries, to calibrate the argument. m is the gross margin for the advertiser's business

We also need to define a few terms to describe the advertiser's budget. Let w be the number of weeks covered by the campaign's analysis (and the advertising ex-

pense), b give the fraction of revenue devoted to advertising (% advertising budget), and R be the total annual revenue. To get the affordability bound, we define γ_C as the fraction of the ad budget in the campaign window devoted to the Super Bowl ad. For instance, if $\gamma_C = 1$, this means the firm spends all advertising dollars for the period in question on the Super Bowl.

We now present the argument, which is an algebraic exercise with one key step: substituting for the coefficient of variation and solving for the revenue bounds.

First we construct the affordability bound. To afford the ad, it must be the case that it costs less than the ad budget, which is the revenue for the time period in question, $R \cdot \frac{w}{52}$, times b, the percentage of the revenue devoted to advertising, times γ_c, the fraction of the budget that can be devoted to one media outlet:

$$C \leq \left(R \cdot \frac{w}{52}\right) \cdot b \cdot \gamma_c.$$

Solving this equation for revenue gives the affordability limit:

$$R \geq \frac{C}{\gamma_C b \cdot \frac{w}{52}}. \tag{11}$$

For the detectability limit, let r and r_0 be the target ROI and null hypothesis ROI, respectively. The t-statistic is given by:

$$\begin{aligned} t_{ROI} &\leq \frac{r - r_0}{\sqrt{\frac{2}{N}} \times \sigma_{ROI}} \\ t_{ROI} &\leq \frac{(r - r_0)}{\sqrt{\frac{2}{N}} \left(\frac{m\sigma}{c}\right)} \\ t_{ROI} &\leq \frac{(r - r_0)}{\sqrt{\frac{2}{N}} \left(\frac{\sigma}{\mu}\right) / \frac{c}{m\mu}}. \end{aligned}$$

The first equation is just the definition of the test statistic. The second equation follows from substituting in the standard deviation of ROI, which is a linear function of the sales standard deviation, per capita cost, and gross margin. The final equation simply multiplies the denominator by $\frac{\mu}{\mu}$. We do this so we can substitute in a constant for the coefficient of variation, $\frac{\sigma}{\mu}$, and solve for μ, as given below:

$$\mu \leq \frac{(r - r_0)\, c}{\sqrt{\frac{2}{N}} \left(\frac{\sigma}{\mu}\right) m \cdot t_{ROI}} \equiv \bar{\mu}$$

The right-most definition is for notational convenience. We can also relate mean sales during the campaign period to total revenue:

$$\mu = R \cdot \frac{\frac{w}{52}}{N_{Total}}. \tag{12}$$

We then solve for revenue and substitute in $\bar{\mu}$ for μ to get the detectability limit:

$$R \leq \frac{N_{Total} \cdot \bar{\mu}}{\frac{w}{52}} \tag{13}$$

Examining the detectability limit, referring back to $\bar{\mu}$ where necessary, we see that it decreases with $\frac{\sigma}{\mu}$. This is intuitive, as the noise to signal ratio increases, inference becomes more difficult. It also falls with the required t and gross margin. To understand why the bound rises as margin falls, consider two companies, one with a high margin, one with a low margin. All else equal, the low margin firm is experiencing a larger change in sales for a given ROI change. Naturally the bound also rises with the gap between the null hypothesis and target ROI.

Putting both limits together, we obtain the interval for detectability and affordability in terms of the firm's annual revenue:

$$\frac{C}{\gamma_C b \cdot \frac{w}{52}} \leq R \leq \frac{N_{Total} \cdot \bar{\mu}}{\frac{w}{52}}. \tag{14}$$

6.2 Targeting details

The standard deviation of the ROI, σ_{ROI}, is given by:

$$\begin{aligned} ROI &= \frac{\Delta\mu(M)}{C(M)} - 1 \\ \sigma^2_{ROI} &= Var\left(\frac{\Delta\mu(M)}{C(M)}\right) = \frac{2\sigma^2(M)}{M \cdot (C(M))^2} \end{aligned}$$

which implies:

$$\sigma_{ROI} = \frac{\sigma(M)}{\sqrt{M/2} \cdot C(M)} \tag{15}$$

Notice that this formula does not rely upon the actual impact of the ads, except that we calibrate the expected effect against the cost (in reality, costs will be correlated with ad impact). It only incorporates the average volatility of the M observations. The standard error of our estimate of the ROI is decreasing in M as long as the ratio $\sigma(M)/C(M)$ does not increase faster than $\sqrt{M}$. For the special case of a constant variance, the standard error of the ROI can be more precisely estimated as long as the average costs do not decline faster than $\frac{1}{\sqrt{M}}$. Note average costs cannot decline faster than $\frac{1}{M}$ unless the advertiser is actually paid to take extra impressions, which seems unlikely. Another special case is constant average cost. Here as long as $\sigma(M)$ does not increase faster than $\sqrt{M}$, more precision is gained by expanding reach.

6.3 Campaign window proof

Note this entire argument is also in a forthcoming NBER book chapter.

We again employ the t-statistic, but also index little t for time. For the sake of concreteness, let time be indexed in terms of weeks. For notational simplicity, we will assume constant variance in the outcome over time, no covariance in outcomes over time,[40] constant variance across exposed and unexposed groups, and balanced group sizes. We will consider the long-term effects by examining a cumulative t-statistic (against the null of no effect) for T weeks rather than a separate statistic for each week. We write the cumulative t-statistic for T weeks as:

$$t_{\Delta\bar{y}_T} = \sqrt{\frac{N}{2}} \left(\frac{\sum_{t=1}^{T} \Delta\bar{y}_t}{\sqrt{T}\hat{\sigma}} \right). \tag{16}$$

At first glance, this t-statistic appears to be a typical $O\left(\sqrt{T}\right)$ asymptotic rate with the numerator being a sum over T ad effects and the denominator growing at a $\sqrt{T}$ rate. This is where economics comes to bear. Since $\Delta\bar{y}_t$ represents the impact of a

[40] This assumption is clearly false: individual heterogeneity and habitual purchase behavior result in serial correlation in purchasing behavior. However, as we are considering the analysis over time, if we assume a panel structure with fixed effect or other residual-variance absorbing techniques to account for the source of this heterogeneity, this assumption should not be a first-order concern.

given advertising campaign during and following the campaign (since $t = 1$ indexes the first week of the campaign), $\Delta\bar{y}_t \geq 0$. But the effect of the ad each week cannot be a constant—if it were, the effect of the campaign would be infinite. Thus it is generally modeled to be decreasing over time.

With a decreasing ad effect, we should still be able to use all of the extra data we gather following the campaign to obtain more statistically significant effects, right? Wrong. Consider the condition necessary for an additional week to increase the t-statistic:

$$\begin{aligned} t_{\Delta\bar{y}_T} &< t_{\Delta\bar{y}_{T+1}} \\ \frac{\sum_{t=1}^{T} \Delta\bar{y}_t}{\sqrt{T}} &< \frac{\sum_{t=1}^{T+1} \Delta\bar{y}_t}{\sqrt{T+1}} \end{aligned}$$

Some additional algebra leads us to

$$1 + \frac{1}{T} < \left(1 + \frac{\Delta\bar{y}_{T+1}}{\sum_{t=1}^{T} \Delta\bar{y}_t}\right)^2$$

which approximately implies

$$\frac{1}{2} \cdot \frac{1}{T} \sum_{t=1}^{T} \Delta\bar{y}_t < \Delta\bar{y}_{T+1}. \tag{17}$$

This last expression says, "If the next week's expected effect is less than one-half the average effect over all previous weeks, then adding it in will only reduce precision." Thus, the marginal week can actually cloud the previous weeks, as its signal-to-noise ratio is not sufficiently large enough to warrant its inclusion.[41] If the expected impact of the campaign following exposure decays rapidly (although not necessarily all the way to zero), it is likely that including additional weeks beyond the campaign weeks will decrease the statistical precision.

Suppose that you were just content with the lower bound of the confidence interval increasing in expectation. A similar calculation, under similar assumptions,

[41] Note that this expression is completely general for independent random draws under any marginal indexing or ordering. In the identically distributed case, though, the expected mean for the marginal draw is equal to all inframarginal draws, so the inequality holds.

shows that the lower bound of a 95% confidence interval will increase if and only if

$$1.96\left(\sqrt{T+1}-\sqrt{T}\right) < \frac{\Delta\bar{y}_{T+1}}{\hat{\sigma}/\sqrt{N}} \tag{18}$$

where the right-hand expression is the marginal expected t-statistic of the $T+1^{th}$ week.

We can summarize these insights by returning to our formula for the t-statistic:

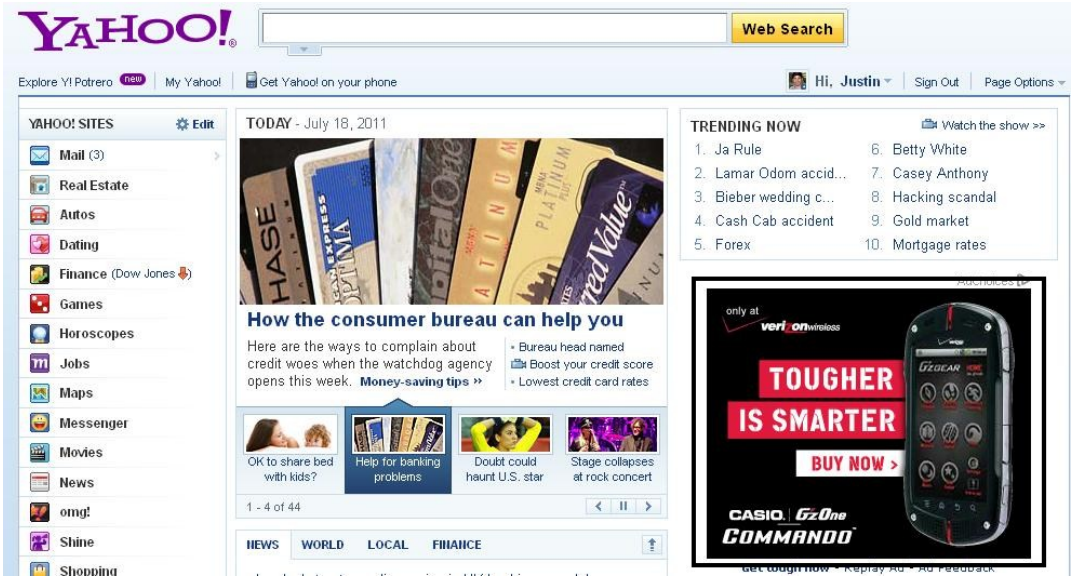
$$t_{\Delta\bar{y}_T} = \sqrt{\frac{N}{2}}\left(\frac{\sum_{t=1}^{T}\Delta\bar{y}_t}{\sqrt{T}\hat{\sigma}}\right).$$

Since the denominator is growing at $O\left(\sqrt{T}\right)$, in order for the t-statistic to grow, the numerator must grow at a faster rate. In the limit we know this cannot be as the total impact of the advertising would diverge faster than even the harmonic series.[42]

Ex-ante it is hard to know when the trade-off turns against you. The effect may decay slower than the harmonic series initially, and then move towards zero quite quickly. Of course if we knew the pattern of decay, we would have answered the question the whole exercise is asking! So in the end the practitioner must make a judgment call. While choosing longer time frames for advertising effectiveness analyses should capture more of the cumulative effect (assuming that it is generally positive), including additional weeks may just cloud the picture by adding more noise than ad impact. Measuring the effects of advertising inherently involves this sort of "judgment call"—an unsatisfying step in the estimation process for any empirical scientist. But the step is necessary since, as we have shown, estimating the long-run effect of advertising is a losing proposition—the noise eventually overwhelms the signal, the question is "when" and right now our judgment call is to use 1–4 weeks, but this is far from the final word.

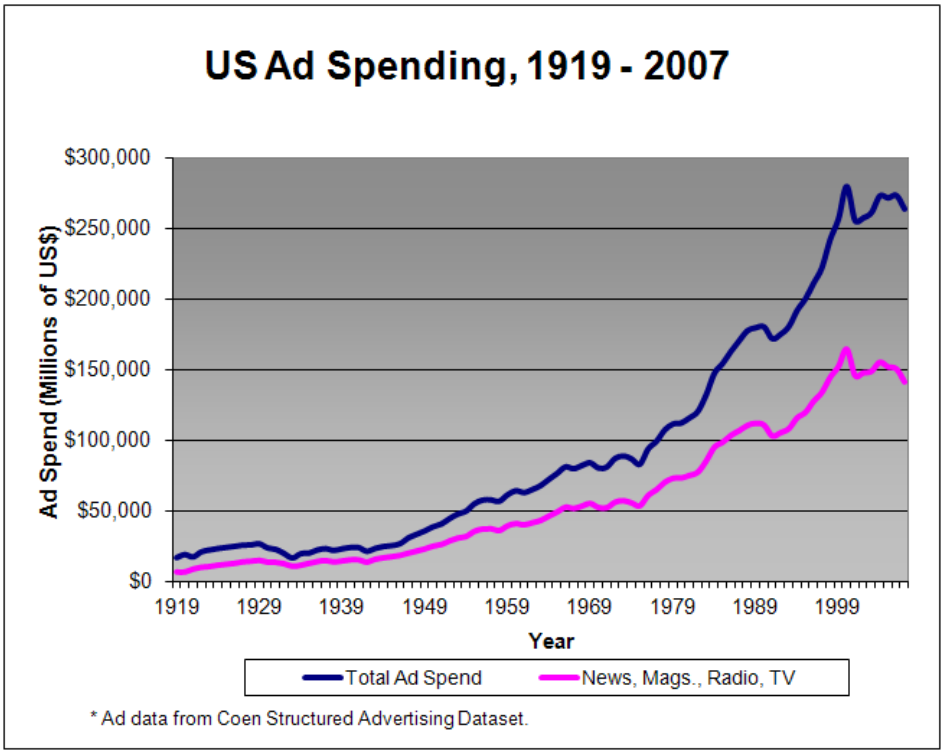
[42]We note that an asset with infinite (nominal) returns is not implausible per se (a consolidated annuity, known as a "consol," does this), but we do find infinite effects of advertising implausible. The harmonic series is $\sum \frac{1}{t}$ whereas the requisite series for an increasing t-statistic would be $\approx \sum \frac{1}{\sqrt{t}}$ which diverges much more quickly.

6.4 Display ad example



Appendix Figure 1: Display ad example on Yahoo.com.

6.5 US ad spending figures



Appendix Figure 2: U.S. Ad Spending 1919–2007.

6.6 Advertising across industries and firms

Appendix Table 1: Advertising Expenditure Across Industries and Firms

Industry/Firm	Revenue In $Billion	Gross margin %	Ad Expenditure In $Billion	Ad Revenue Share %
Mobile Carriers				
Verizon	114.2	56.9%	1.56344	1.37%
Sprint Nextel	35.1	41.8%	0.67308	1.92%
ATT	127.4	54.5%	1.73602	1.36%
T-Mobile	19.2	N/A	0.52627	2.75%
Automakers				
Honda	115.1	21.4%	0.57124	0.50%
Toyota	262.2	10.2%	0.85032	0.32%
Ford	133.3	17.2%	0.87670	0.66%
GMC	150.1	12.7%	0.17907	0.12%
Fiat-Chrysler	55.0	5.5%	0.87490	1.59%
Hyundai	74.0	N/A	0.30144	0.41%
Dodge	N/A	N/A	0.52501	N/A
Rental Cars				
Avis Budget Group	6.7	24.5%	0.04520	0.67%
Hertz	8.6	43.2%	0.03735	0.43%
Enterprise/Alamo	13.5	N/A	0.06733	0.50%
Dollar Thrifty	1.5	33.7%	0.00021	0.01%
Airlines				
American (AMR)	24.9	47.4%	0.06034	0.24%
United	37.4	56.3%	0.03313	0.09%
Delta	36.5	39.0%	0.05801	0.16%
US Airways	13.7	33.9%	0.01151	0.08%
Online Brokerages				
Scottrade	0.8	100.0%	0.07084	8.45%
Etrade	1.3	100.0%	0.16672	12.63%
TD Ameritrade	2.8	100.0%	0.05034	1.82%
Fast Food				
McDonald's	27.4	39.0%	0.95926	3.50%
Burger King	2.3	37.6%	0.29712	12.92%
Wendy's	2.4	25.3%	0.27248	11.21%
Dairy Queen	2.5	N/A	0.07276	2.91%
Jack in the Box	2.2	45.2%	0.07253	3.30%