The Unfavorable Economics of Measuring the Returns to Advertising*

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

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

September 18, 2014

Abstract

Twenty-five large field experiments with major U.S. retailers and brokerages, each reaching millions of customers and collectively representing $2.8 million in advertising expenditure, reveal that measuring the returns to advertising is exceedingly difficult. The median confidence interval on ROI is over 100% wide, the smallest exceeds 50%. Detailed sales data show that, relative to the per capita cost of the advertising, individual-level sales are incredibly volatile; a coefficient of variation of 10 is common. Hence, informative advertising experiments can easily require more than ten million person-weeks, making experiments costly and potentially infeasible for many firms. Despite these unfavorable economics, randomized control trials represent progress by injecting new, unbiased information into the market. The statistically small impact of profitable advertising amid such noise means that selection bias is a crippling concern for widely-employed observational methods. We discuss how these biases and weak informational feedback from experiments fundamentally impact both advertisers and publishers.

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 the Returns to Advertising." We especially thank David Reiley for his contributions to this work. Ned Augenblick, Arun Chandrasekhar, Sharad Goel, Garrett Johnson, Clara Lewis, R. Preston McAfee, Markus Möbius, Lars Lefgren, Michael Schwarz and Ken Wilbur 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. The work represents our views and not those of our current or former employers.

1 Introduction

Each day, a typical American views 25–45 minutes of television commercials, many billboards and numerous online ads amounting to an annual U.S. advertising expenditure of roughly $500 per person (Kantar Media, 2008). To break even, the universe of advertisers needs to net $1.50 in per capita profits per day, which corresponds to $4–6 in incremental sales (revenue) or about $3,500–5,500 annually per household. So while a large impact is required for advertising to be profitable in aggregate it is very much an open question as to whether advertising causally impacts consumers to this degree. This uncertainty is perfectly captured by the *first sentence* of an influential paper on advertising effectiveness "Until recently, believing in the effectiveness of advertising and promotion was largely a matter of faith" (Abraham and Lodish, 1990).[1] In this paper we address the underlying puzzle: if so much money is being spent on advertising, how could it be possible that firms have such imprecise beliefs on its returns?

At the heart of our analysis is a data-driven argument demonstrating that the economics of measuring the returns to advertising are very unfavorable and while advances in measurement technology are improving, the baseline we measure indicates that imprecise beliefs on advertising effectiveness are likely to persist for the foreseeable future. Our findings center around 25 large-scale digital advertising *field experiments*—accounting for $2.8 million in expenditure—from well-known retailers (19) and financial service firms (6) partnering with a large web publisher. Digital advertising allows for ad delivery and consumer purchases to be measured at the individual level and randomized to ensure exogenous exposure, but our findings are not restricted to this particular delivery channel.[2] Our results are stated in dollar terms: the causal sales impact based on per-person expenditure. Presenting results in dollar space in turn allows us to draw useful comparisons to media such as television, where emerging technologies afford randomized and personalized ad delivery. Finally, while

[1] Before this sentence was written, American firms had spent approximately $4.6 trillion promoting their products and services (Coen Structured Advertising Dataset, 2007). The authors pioneering field experiments with split-feed cable TV produced some of the first evidence that advertising causally influenced consumers.

[2] For many media, geo-randomized advertising experiments are state of the art. Experiments that rely on this method are significantly more expensive because you are unable to eliminate the noise from purchases among those who the advertiser is unable to reach with their message (similar to issues surrounding intent-to-treat).

we believe our central findings apply to most advertising dollars and advertising media, we are careful to discuss delivery channels and advertiser segments where they are unlikely to hold.

We start with a simple model of the firm's advertising problem. The key parameter is return on investment (ROI)—the profits generated through advertising as a percentage of cost. ROI captures what the firm presumably cares about in the end, profits, and our extensive data sharing agreements allow us to sidestep intermediate metrics such as clicks.[3] For retailers, the data comes from online and offline sales tracking. For financial firms, we convert account sign-ups into a dollar figure based on the average lifetime value of a customer. These purchase data measured at the individual level—some of the first of their kind to end up in published work—reveal the source of the inference difficulty. The standard deviation of individual-level sales is typically *ten times* the mean over typical campaign duration and post-campaign evaluation window. While this relationship may not hold true for smaller firms or new products, it was remarkably consistent across the relatively diverse set of advertisers in our study.

Each experiment 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—those targeted by the campaign—from receiving an advertiser's online display ad. Campaign costs ranged from \$0.02–0.35 per exposed user, most were close to \$0.10. This corresponds to 20–60 "premium" display ads or about 7–10 prime-time television commercials.[4] We convert the sales impact to profits by using gross margins reported to us by the partner firms and checked against SEC filings.[5] The median standard error on ROI is 26.1%, implying a confidence interval over one hundred percentage points wide. For advertising at this level of intensity, massive trials, typically in the single-digit millions of person-weeks are required to reliably evaluate a pure waste of money (-100% ROI) from profitability (ROI>0%).

To provide more depth and in the interest of expanding the applicability of our

[3]Further on we discuss cases in which intermediate metrics can be used with limited induced bias, however generally there are serious complications that can arise from using these metrics, discussed in more detail in Lewis, Rao and Reiley (In Press).

[4]Due to greater targeting variation, even online ads in desirable spots have more widely varying prices than other media.

[5]We used a single margin for each firm. By using a constant margin we are assuming away some of the uncertainty surrounding measuring returns, as many firms have widely varying margins in their product lineups.

findings we compute how large each experiment would have to be to reliably evaluate various hypothesis sets of interest. In this thought experiment, we assume the advertiser could add new, independent person-weeks to their campaign.[6] The median campaign would have to be nine times larger to reliably distinguish between a wildly profitable campaign (+50% ROI) from one that broke even (0% ROI). Using tolerances more commonly used for investment decisions, such as a 10% ROI difference, requires the median campaign to be *62 times larger* (mean 421) to possess adequate power—nearly impossible for a campaign of any realistic size. And while ROI measures *average* returns, we briefly discuss the (rather incredible difficulties) in determining the average ROI target that corresponds to zero marginal profit.

While our analysis shows that even very large randomized control trials (RCTs) can be surprisingly uninformative, RCTs are nonetheless a huge step forward in the science and measurement of advertising. In particular, our experimental data suggest that very large biases lurk undetected in commonly-employed observational methods (e.g., one such bias was uncovered by experiments in (Lewis et al., 2011)). These biases exist primarily because ads are, entirely by design, not delivered randomly—a marketer's job is to target campaigns across consumers, time and context. Suppose we evaluate a campaign with a regression of sales per individual (in dollars) on an indicator variable of whether or not the person saw a firm's ad. In an experiment, the indicator variable is totally exogenous, while in an observational method, one attempts to neutralize selection bias induced by targeting. To net a +25% ROI, our median campaign had to causally raise average per person sales by \$0.35. Calibrating this against sales volatility, the goal is to detect a \$0.35 impact on a variable with a mean of \$7 and a standard deviation of \$75. In terms of model fit, the R^2 for a *highly profitable* campaign would be on the order of 0.0000054.[7] To successfully employ an observational method, one must not omit endogenous factors or misspecify functional form to a degree that would generate an R^2 on the order of *only 0.000001*. This appears to be an impossible statistical feat in an environment where selection effects are expected to be as large as 30 times the true treatment effect.[8]

[6] In practice they may also opt to run a more concentrated test. A larger per-person spend makes the inference problem easier, but advertising at undesirably high intensity can bias the measured effect.

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

[8] To see the size of selection effects, consider a simple example: if a campaign spends 10 cents per individual (20–40 "premium" display ads or about 10 prime-time television commercials) and

Since we are making the (admittedly) strong claim that most advertisers do not, and indeed some *cannot*, know the effectiveness of their advertising spend, it is paramount to investigate the generalizability of our empirical findings. First, the retail and financial services firms we study are representative in terms of revenue base, margins, and product types of firms that constitute the majority of ad spending. Second, we show that holding expenditure fixed and lengthening the evaluation window would typically not improve statistical power. Third, we made our best effort with all the data at our disposal, such as pre-campaign sales, to control for factors that may have differed by chance between the treatment and control group, thus improving power. Fourth, our experimental size multipliers help calibrate the financial commitment necessary to produce truly informative RCTs and thus can be used to evaluate the applicability of our findings to specific firms or market segments.

Despite the fact that we believe our results are broadly applicable, a few important caveats are worth mentioning. Our results are unlikely to apply to firms or products with low baseline sales volatility, such as firms who receive nearly all their customers from advertising (*e.g.* direct-response TV ads) and products for which there is no baseline awareness. Further, our results show that the outlook is more positive for campaigns that all else equal have substantially higher per-person expenditure.[9]

The unfavorable economics of measuring the returns to advertising have several important implications. First, scarce information means there is little "selective pressure" on advertising levels across firms. With supplemental data we examine several major industries and find that otherwise similar firms often have vastly differing levels of advertising expenditure. Second, if experimentation becomes more common, consistent with trends in experimentation technology, massive publishers will be conferred an additional strategic advantage of scale, given the size of RCTs required to provide reliable feedback on ROI. Third, imprecise signals on the returns to advertising introduce issues involving strategic misreporting. Finally, we note that while there are certainly other investments and decisions that firms make that have hard-to-measure

consumers have unit-demand for a product that returns marginal profit of $30, then only 1 in 300 people need to be "converted" for the campaign to break even. Suppose a targeted individual has a 10 percentage points higher baseline purchase probability (a realistic degree of targeting similar in magnitude to Lewis and Reiley (2014)), then the selection effect is expected to be *30 times larger* than the causal effect of the ad.

[9]Concentrating expenditure can increase power; however, diminishing returns may imply that precisely when one can measure the returns, an economically unfavorable (i.e., unprofitable) result is guaranteed.

returns, such as management consulting (Bloom et al., 2013), capacity expansion or mergers, but what differs here is that the metrics and methods used in the advertising industry produce a veneer of quantitative certitude that is not typically found in these other circumstances. Using RCTs to reveal the true uncertainty in measuring returns and, in the process, exposing non-experimental techniques consequently has very different implications for the advertising market.

2 The Advertiser's Problem

In this section we formalize and calibrate the problem of campaign evaluation.

2.1 Definitions and model

We define a campaign as a set of advertisements delivered to a set of consumers through a single channel over a specified (and typically short) period of time using one "creative" (all messaging content such as pictures, text, and audio). Ex-post evaluation asks the question, "Given a certain expenditure and delivery of ads, what is the rate of return on this investment (ROI)? Side-stepping broader optimization issues, we take the target population as given and focus on measurement of the return on investment.

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 is defined by a continuous concave function of per-user expenditure $\beta(c)$.[10] We 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 that $\beta(c) * m$ gives gross profit per person. Net profit subtracts cost $\beta(c) * m - c$, and ROI measures net profit as a percentage of cost $\frac{\beta(c)*m-c}{c}$. In our simple model the only choice variable is c, or "how much I advertise to each consumer."

Figure 1 graphically depicts the model: c^* gives optimal spend and c_h gives the spend level where ROI is exactly 0%. At any point past c_h the firm has negative

[10]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.

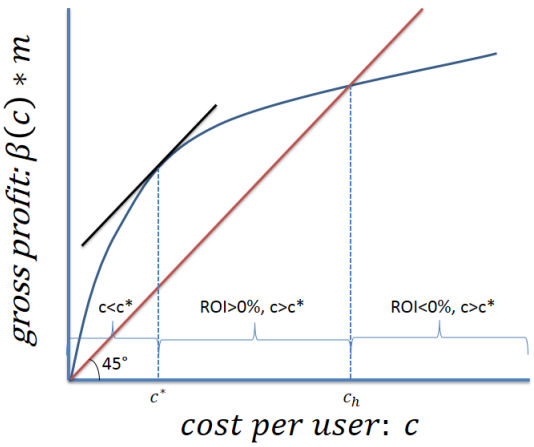


Figure 1: Graphical depiction of the advertiser's problem.

returns, whereas any point to the left of c^* the firm has positive returns but 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 and the actual quantity of ads delivered per person is stochastic (because exposure depends on user activity). Our 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 prominence is also consistent with our personal experience in the industry.

2.2 Calibrating the model with data

We now calibrate the campaign evaluation with data from experiments. On the cost side, display ad campaigns that deliver a few ads to each exposed user per day cost about 1–2 cents per person per day and typically run for about two weeks, cumulating in a cost between 10 and 40 cents per person. This corresponds to about one 30-second TV ad per person per day. Given the total volume of advertising, a typical consumer sees across all media, even an intense campaign only captures about 2% of a user's

advertising "attention.

As mentioned in the introduction, a key source of the inference challenge facing advertisers is sales volatility. For a given individual, it has three components: probability of making a purchase, basket size conditional on purchasing and frequency of purchases. For an advertiser, these components vary by user and all contribute to total sales volatility. For the large retailers and financial service firms in our study, mean weekly sales per-person varies 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) is far more uniform. For retailers, it ranges from 4–23, but tends to be clustered between 10–15. Customers buy goods relatively infrequently, but when they do, transaction values are quite volatile about the mean. For the financial service firms, we assume a uniform lifetime value for each new account acquired as a result of advertising. While this assumption eliminates the basket-size component of sales variance, financial firms still face a considerably higher coefficient of variation because new accounts are rare and the lifetime value tends to be quite high. Automobiles and other big ticket items also share this feature.[11]

Let y_i give sales for individual i. We assume, for simplicity, that each affected individual saw the same value of advertising for a given campaign, so let indicator variable x_i quantify ad exposure. $\hat{\beta(c)}$ gives our estimate of the sales impact 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 notation would use $y|W$. However, even in an experiment, power can be improved by conditioning on exogenous and predetermined factors, such as pre-campaign sales, that are predictive of baseline purchases or that may have differed by chance between treatment and control subjects. We suppress the additional covariates below for this illustrative example, but we stress that they are in fact used in our empirical specifications to soak up residual variation. Hence, all the following results are qualitatively unaffected by such modeling improvements up to the usual "conditional upon" caveat where the R^2 becomes partial R^2 of the treatment variable.

[11] In contrast, homogeneous food stuffs have more stable expenditure, but their very homogeneity likely reduces own-firm returns to equilibrium levels of advertising within industry as a result of positive advertising spillovers to competitor firms (Kaiser, 2005).

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

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

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—for a thorough explanation of this algebra see Lovell, 2008): :

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

In this model, R^2 can be usefully expressed as a function of ratio of the sales difference between exposed and unexposed groups and the standard deviation in sales. We can express the t-statistic as for testing the hypothesis ($\beta = 0$) as a function of this ratio

as well:

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

We call $(\frac{\Delta\bar{y}}{\sigma})$ the "impact-to-standard-deviation ratio."[12]

We calibrate these statistics using a representative experiment—slightly larger than the median—from our study. For ease of exposition, we will discuss the hypothetical case as if it were a single, actual experiment. The cost per exposed user is \$0.14 (20–80 display ads or 7–10 TV commercials) and gross margin is 50%. Mean sales per-person for the period under study is \$7 with a standard deviation \$75.

We suppose the ROI target is 25%, which, given margins, corresponds to a \$0.35 sales impact per person. A \$0.35 per-person impact on sales is a 5% increase in sales during the two weeks of the campaign. In terms of percentage lift, the required impact of the campaign appears quite large. 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.[13] From our derivation above, this implies an R^2 of:

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

Perhaps surprisingly, even a very successful campaign has a minuscule R^2 of *0.0000054*.[14] An immediate consequence is that a very large N is required 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 with a null hypothesis of -100% ROI (zero causal impact) is 3.30. 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 alterna-

[12] It is also known as *Cohen's d*.

[13] This is less than 1/40th the "Small" effect size of 0.2 outlined in Cohen (1977).

[14] This is less than 1/10,000th the R^2 of $\approx$ 6% in the low-powered examples for advertising's impact on sales in the discussion of statistical power in marketing in Sawyer and Ball (1981), though their examples reflect aggregate store-level, rather than customer-level models.

tive hypothesis of a healthy 25% ROI, we fail to reject the null that the ad had no causal impact 74% of the time.[15]

The tiny R^2 for the treatment variable not only reveals the unfavorable power of RCTs, but 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 intertemporal correlation between ad exposure (web browsing) and shopping (Lewis et al., 2011) 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 of wage/schooling regressions, in which the endogeneity has often been found to be 10-30% of the treatment effect (Card, 1999). A minimal level of targeting that results in the exposed group having only a few percentage points higher baseline purchase rate can lead to an expected bias many multiples of the treatment effect. Unless this difference is controlled for with near *perfect* precision, observational models will have a large bias.

It may appear that observational models are so ill-suited for this setting that we are arguing against a straw man, but these techniques are commonly employed in the industry. A relatively recent *Harvard Business Review* article co-authored by the president of comScore, one of the largest data-providers for web publishers and advertisers, reported a 300% impact of online advertising (Abraham, 2008). Their estimate is generated from a regression-based comparison of endogenously exposed and unexposed groups. This estimate seems surprisingly high as it implies that advertising prices should be at least an order of magnitude higher than current levels. The use of these techniques in industry is also discussed in the experimental work of (Blake et al., 2013).

3 Analysis of the 25 Field Experiments

In this section we delve into the field experiments. There is an inherent challenge in discussing this many experiments in adequate detail. We put detailed information in two comprehensive summary tables and limit our discussion in the text to more

[15] 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 estimated effect is positive. This overestimation was recently dubbed the "exaggeration factor" by Gelman and Carlin (2013).

general points. We limit our reporting to the statistical uncertainty surrounding the measurement of advertising returns and do not report the point estimates for each campaign. This is for a few reasons. First, while we can cite evidence that these firms are representative of a typical advertising dollar, we cannot cite evidence that they are representative of typical effectiveness. Second, reporting imprecisely estimated means could potentially be misleading, due to the "exaggeration factor" from low-powered tests (Gelman and Carlin, 2013). Third, confidentiality agreements limit us from sharing all the point estimates, which raise concerns about selection effects for the ones we can report.

3.1 Overview and data description

Table 1 gives an overview of 25 display advertising experiments/campaigns. Due to confidentiality agreements, we cannot reveal the identity of the advertisers. They are large retailers (Panel 1) and financial service firms (Panel 2) that are most likely familiar to American readers. We employ a naming convention using the vertical sector of the advertiser in lieu of the actual firm names; these are given in Column 1 and the year of the experiment is given in Column 2.[16]

Sales is the key dependent measure for the firms in Panel 1, and Column 3 gives the unit of observation: for ten experiments the data is daily (indicated by "1"), five it is weekly ("2") and four it is the campaign period ("3"). In Panel 2, the dependent measure is new account sign-ups, which was collected over the campaign period except in one case, where it was observed daily. Columns 5 and 6 provide institutional details of the experiments that may be of interest to readers familiar with online experimentation. Column 8 shows that the experiments ranged from 2 to 135 days, with a median of 14 days. Column 9 shows that campaign cost varied from relatively small ($9,964) to quite large ($612,693). The mean was $114,083; the median was $75,000. The median campaign reached over one million individuals, and all campaigns had hundreds of thousands of individuals in both test and control cells (as shown in Columns 9–11). Combined, the campaigns represent over $2.8 million in expenditure.

Column 7 gives the control variables we have available to reduce noise in the

[16]Many of the experiments are taken from past work from Yahoo! Labs, like Lewis and Reiley (2013); Lewis and Schreiner (2010); Johnson, Lewis and Reiley (2014); and Lewis, Rao and Reiley (2011).

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	2	1	-	1,2,3	14	$128,750	1,257,756	300,000	814,052	-	$9.49	$94.28
R 1	2007	2	1	-	1,2,3	10	$40,234	1,257,756	300,000	686,878	-	$10.50	$111.15
R 1	2007	2	1	-	1,2,3	10	$68,398	1,257,756	300,000	801,174	-	$4.86	$69.98
R 1	2008	2	1,2,3	-	1,2,3	105	$260,000	957,706	300,000	764,235	238,904	$125.74	$490.28
R 1	2010	2	1,2	-	1,2,3	7	$8433	2,535,491	300,000	1,159,100	-	$11.47	$111.37
R 1	2010	1	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	3	1	-	-	35	$191,750	3,145,790	3,146,420	2,229,959	-	$30.77	$147.37
R 2	2009	3	1	-	-	35	$191,750	3,146,347	3,146,420	2,258,672	-	$30.77	$147.37
R 2	2009	3	1	-	-	35	$191,750	3,145,996	3,146,420	2,245,196	-	$30.77	$147.37
R 3	2010	1	1,2,3	1	1,3	3	$9,964	281,802	161,163	281,802	161,163	$1.27	$18.46
R 3	2010	1	1,2	1	1,3	4	$16,549	483,015	277,751	424,380	-	$1.08	$14.73
R 3	2010	1	1,2,3	1	1,3	2	$25,571	292,459	169,024	292,459	169,024	$1.89	$18.89
R 3	2010	1	1,2,3	1	1,3	3	$18,234	31566	179,709	31566	179,709	$1.29	$16.27
R 3	2010	1	1,2	1	1,3	3	$18,042	259,903	452,983	259,903	-	$1.75	$18.60
R 3	2010	1	1,2,3	1	1,3	4	$27,342	355,474	204,034	355,474	204,034	$2.64	$21.60
R 3	2010	1	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	1,2	1	1	18	$90,000	1,075,828	1,075,827	693,459	-	$0.56	$12.65
R 5	2010	3	1,2	-	1,3	41	$180,000	2,321,606	244,432	583,991	-	$54.77	$170.41
R 5	2011	1	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			New	Pr New	SE New
								Test	Control	Exposed	Accts	(Test)	Acct
F 1	2008	3	1,2,4	-	3	42	$50,000	12% of Y!	52% of Y!	794,332	867	0.0011	0.0330
F 1	2008	3	1,2,4	-	3	42	$50,000	12% of Y!	52% of Y!	748,730	762	0.0010	0.0319
F 1	2008	3	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	3	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	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	3	1,2	-	3	36	$85,942	8,125,910	8,125,909	793,042	1090	0.0014	0.0331

* Estimation strategies employed to obtain the standard errors of the ad impact between the test and control groups follow:
"Y" 1:Daily, 2:Weekly, 3: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.

experimental estimates. These include lagged sales (indicated by "1"), demographics ("2") and online behaviors (such as intensity of browsing) "3", all measured at the user level. Lagged sales were available for 16 of the 19 retail experiments, but are not used for the financial service firms since these campaigns were designed to produce new account sign-ups. Demographics were available for six experiments and online behavior was available for all but Retailer 2. We used the appropriate panel techniques to predict and absorb residual variation. Lagged sales are the best predictor, reducing variance in the dependent variable by as much as 40%. These reductions (and the importance of lagged sales) are consistent with related work (Deng et al., 2013). A little math shows that going from $R^2 = 0$ in the univariate regression to $R^2_{|\mathbf{W}} = 0.40$ yields a sublinear reduction in standard errors of 23%.[17] An order-of-magnitude reduction in standard errors would require $R^2_{|\mathbf{W}} = 0.99$.

The second-to-last column gives the average sales per customer. It varies quite widely across retailers, which is due to differing firm popularity and the degree of targeting used in the campaign (a more targeted campaign typically has higher baseline sales). Across experiments, median per person sales 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 its mean per-person sales. The standard-deviation-to-mean ratio exceeds seven for all but two experiments. Longer campaigns tend to have lower noise-to-signal ratios, which is due to sufficient independence in sales across weeks, but in estimation of shorter experiments some of these efficiency gains can be had by conditioning on pre-period sales.[18]

3.2 Estimating ROI

We start by defining evaluation windows, a necessary step in any empirical analysis of advertising. In working with our partner firms, we followed standard industry practice of including the campaign period and typically a relatively short window of 1–4 weeks following the campaign.

[17] $1 - \sqrt{\frac{1-R^2_{|\mathbf{W}}}{1-R^2}} = 1 - \sqrt{1 - R^2_{|\mathbf{W}}} = 23\%$

[18] 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. Further, while longer campaigns generate more points of observation, these additional data will only make inference easier if the spending per person per week is not diluted.

In principle, the effects of advertising could be very long-lived and therefore the bias minimizing window would be correspondingly long. It turns out, however, that long windows substantially damage power. In Lewis et al. (In Press), we establish the following condition: if the next week's expected impact is less than one-half the average impact over all previous weeks, then including it reduces the t-statistic of the total treatment effect. The proposition tells us when a marginal week introduces more noise than signal and thus hurts estimation precision.[19] This implies that unless there is a very limited decay in the ad effect over time, short windows are optimal from a power perspective.[20] Our focus on the short-run, thus, should be viewed as making the problem as easy as possible. And while there is evidence that marketing can induce buy it now or never" purchasing patterns (Damgaard and Gravert, 2014), when the impact does not occur soon after ad exposure we will in general overestimate statistical precision by using the relatively short chosen evaluation windows.

With our evaluation windows in hand, 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 conditional on the control variables outlined in Column 7 of Table 1 in order to obtain as precise an estimate as possible. In Column 4, we translate this into 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 ranges from \$0.02–0.39 and is centered around \$0.10. Per person expenditure can be compared to the standard error of the treatment effect given in Column 3 to show how the magnitude of statistical uncertainty in sales relates to expenditure. This figure provides a useful benchmark not only for online advertising expenditures but also for media purchases in other advertising channels.

In Column 7 we translate the sales impact standard errors to ROI using our estimates of gross margins given in Column 6. In Panel 2, we convert a customer acquisition into a fixed dollar value representing the lifetime value of a customer. This ignores any impact on the intensive margin, since the advertising is restricted from affecting potentially heterogeneous value extracted over the lifespan of an acquired

[19] 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 $\frac{5+2}{2}/2 = 1.75$.

[20] The proposition provides helpful guidance and explains why short windows are used in practice, but quantitatively applying it requires precise ROI estimates for the very inference problem we are trying to solve.

customer. This assumption can be problematic. Just as sales variability has three components, so does the treatment effect on returns. For a given individual, the advertising may cause an increase in the probability of making a purchase, the basket size conditional on purchasing or the frequency of purchases. While we may wish to reduce the treatment effect variability by restricting one or more of these components to be zero, doing so naturally induces a bias-variance trade-off. For example, if advertising only affects the likelihood of purchase, then we will do well to eliminate the variance introduced by basket size and purchase frequency by converting sales into a binary variable. However, if advertising also affects basket size (Lewis and Reiley, 2014) or purchase frequency (Johnson et al., 2014), then we would induce downward bias in the estimate of ROI. In contrast to the evaluation window trade-off where some temporal decay of the treatment effect is natural to expect, it is generally unclear ex-ante which components of sales impact a campaign will primarily influence. Indeed, in the two cited papers, the authors would have failed to reject the null hypothesis of -100% ROI if sales had been converted to a binary variable, but do indeed find significant effects with the continuous measure even though the coefficient of variation was up to 40% lower in the restricted case. Fortunately, for the retail firms we can appropriately account for all the variability in treatment effect when estimating an advertiser's ROI.

Across the retail experiments, the median standard error for ROI is 26.1% (mean 61.8%), implying that the median confidence interval is about *100 percentage points wide*, too wide to be of much practical use. The financial service firm experiments had lower per person expenditure, median \$0.065, and high sales variation (given the all-or-nothing customer acquisition problem) and accordingly had higher standard errors on ROI, with all but one campaign exceeding 93%.

In Figure 2 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. We have already discussed why this is the case for financial services; Retailer 1 simply had a higher standard deviation of sales than 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 campaign requires a larger impact on sales to deliver

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

In-Store + Online Sales

		Key Statistical Properties of Campaign					Ads did anything? H0: ROI=-100% Ha: ROI=0%		Highly profitable? H0: ROI=0% Ha: ROI=50%		Strong performer? H0: ROI=0% Ha: ROI=10%		Maximized profits? 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					Ads did anything? H0: ROI=-100% Ha: ROI=0%		Highly profitable? H0: ROI=0% Ha: ROI=50%		Strong performer? H0: ROI=0% Ha: ROI=10%		Maximized profits? 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

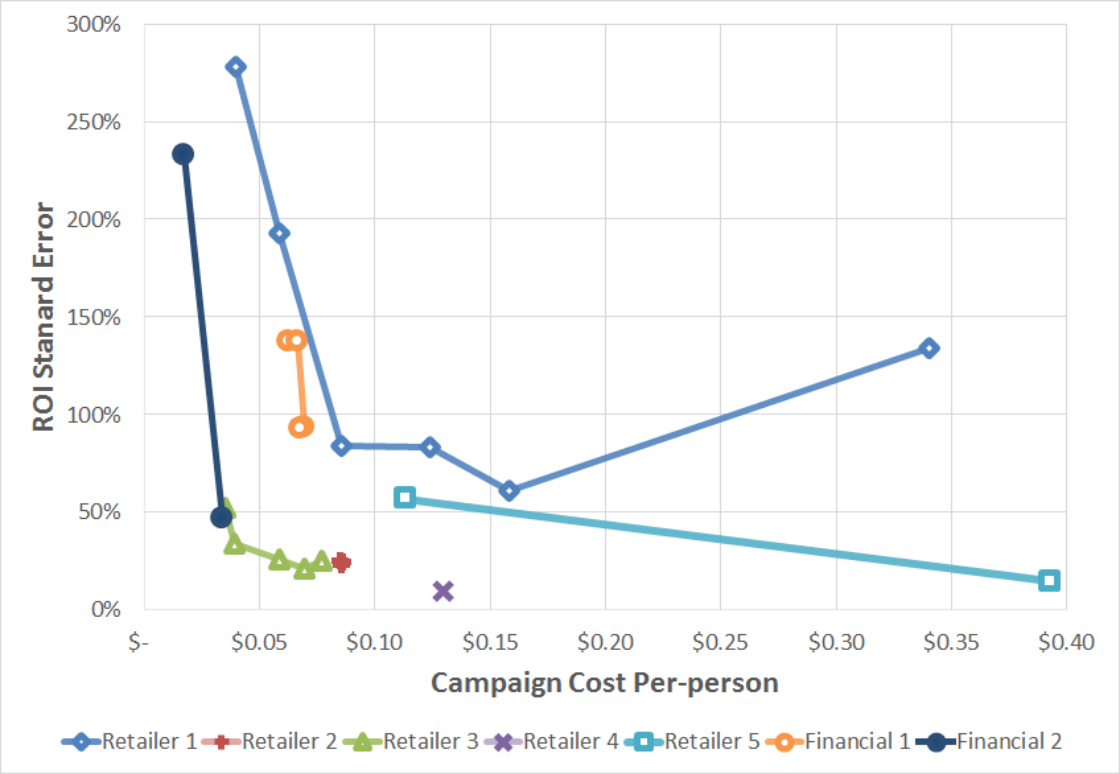


Figure 2: Relationship between ROI uncertainty and campaign cost.

the same percentage return. Measured against the same background noise, a larger impact is easier to identify than a smaller one—the more intense the experiment, the better the power. The most intense spend was Retailer 5's second experiment, at \$0.39 per exposed person, which corresponds to a huge volume of ads: over 80 display ads or 30–40 TV commercials. This experiment also had a large, 3.5 million person control group and pre-period sales and online behaviors to condition on. Yet despite these design advantages, the 95% confidence interval on ROI is still 60 percentage points wide.

Figure 2 highlights a few important points about firm heterogeneity. If more intense spending is likely to be efficient, then it will be easier to evaluate these types of highly-targeted campaigns. We discuss later how this approach could be taken to evaluate effectiveness even if the firm would not use such intense campaigns in normal practice—this approach is yet another bias-variance trade-off that can be made. Conversely, for firms for which low spending per person is likely to be optimal,

then measuring returns will be substantially more difficult.

In the final eight columns of Table 2, we examine each advertiser's ability to evaluate various sets of hypotheses on the returns to expenditure. We start with disparate null and alternative hypotheses and then draw the hypotheses closer to tolerances more typical of investment decisions. For each hypothesis set, we give the expected t-statistic, $E[t]$, to reject the null hypothesis, which is 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%.

For each hypothesis set, we also give an "experiment multiplier," which tells us how much larger the experiment (and implicitly the total cost) would have to be in terms of new, independent individuals 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 expenditure per week. Here we focus on N because it does not require us to model the within-person serial correlation of purchases or the impact function of longer exposure duration. Naturally, if individuals' purchases were independent across weeks and the ad effect was linear, then adding a person-week could be done just as effectively by adding another week to the existing set of targeted individuals.[21] The multipliers are a crucial component of our analysis because they get to the heart of the economics of measuring the returns to advertising. They define the the financial commitment necessary to generate reliable feedback regarding ROI, provided expanding the campaign to the specified degree was indeed possible. They also help extend our results to other media or firms, where larger experimentation may be possible.

We start with distinguishing no impact (-100% ROI) from positive returns (ROI> 0%). Indeed most papers on ad effectiveness use this is as the primary hypothesis of interest—the goal being to measure whether the causal influence on sales is significantly different from zero (Bagwell, 2005).[22] Nine experiments had $E[t] < 1.65$ (Column 9), meaning the most likely outcome was failing to reject -100% ROI when

[21] If the serial correlation is large and positive (negative), then adding more weeks is much less (more) effective than adding more people. Note also that campaigns are typically short because firms like to rotate the creative so that ads do not get stale and ignored.

[22] Specific examples using field experiments include estimating the impact of enlistment recruiting (Carroll et al., 1985), TV commercials for retailers (Lodish et al., 1995; Joo et al., 2013), various media for packaged foods (Eastlack Jr and Rao, 1989) and online ads (Lewis and Schreiner, 2010; Lewis et al., 2011)

the truth was the ad was profitable.[23] Ten experiments had $E[t] > 3$, meaning they possessed sufficient power to reliably determine if the ads had a causal effect on consumer behavior. The remaining six were moderately under-powered.

Simply rejecting that a campaign was a total waste of money is not an ambitious goal. In the next hypothesis we set the null hypothesis as ROI=0% and the alternative to a blockbuster return of 50% (although one could think of this as rejecting a substantial loss -25% in favor of a very strong gain +25%, an ROI difference of 50%). Here twelve experiments had $E[t] < 1$ (severely under-powered), four had $E[t] \in [1, 2]$, five had $E[t] \in [2, 3]$ (90%>power>50%), and only three had $E[t] > 3$. Thus, only three of the twenty-five had sufficient power to reliably conclude that a *wildly profitable* campaign was worth the money, and an additional five could reach this mark by increasing the size of the experiment by a factor of about 2.5 (those with $E[t] \in [2, 3]$) or by using other methods to optimize the experimental design. The median campaign would have to be *nine* 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 being 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 2. 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 an 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 involving millions of individuals. We next draw the hypotheses to a more standard tolerance of ten percentage points, noting that while we use 0% and 10% for instructive purposes, in reality the ROI goal would need to be estimated as well (we discuss this later). Every experiment is *severely* under-powered 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 seven times larger to have sufficient power to distinguish this difference. The median retail sales experiment would have to be *61 times larger*

[23] 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.

(with nine exceeding 100x) 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 financial service firms the median multiplier is a woeful *1241*.

In the final two columns of Table 2, we push the envelope further, setting the difference between the test hypotheses to five percentage points. The multipliers 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. In fact, 5% ROI in our setting is for roughly a two-week period, which corresponds to an annualized ROI of over 100%. If we instead focused on 5% *annualized* ROI, the problem would be 676 times harder.[24]

We note that 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 patent-protected 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, and while this does guard against the idiosyncratic risk of a "dud" campaign, it does not guarantee the firm is at a profitable point on the β function because ex-post measurement is so difficult. Other revenue generating factors of production, such as management consulting, capacity expansion and mergers may involve similar statistical uncertainty. 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 report a positive effect of consulting but also report that precise ROI statements are incredibly difficult to make. A key difference is that the advertising industry is replete metrics and analysis that offer a deceptive quantitative veneer—one might have thought that the ability to randomize over millions of users would naturally lead to precise estimates, but this is not the case for a large share of advertising spend. Observational methods claiming to do better than the levels of efficiency we report (conditional on sample size, etc.) should be viewed with extreme skepticism.

[24] We are trying to estimate 1/26th of the previous effect size, which is 26^2 times harder.

3.3 Determining the ROI target

Returning to Figure 1, there are three important regions separated 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. In this region, although ROI is positive, spending should be reduced, but may interact with the decision-maker's average-marginal bias (de Bartolome, 1995). When $c > c_h$, ROI is negative, and the plan of action is much clearer.

A seemingly straightforward strategy to estimate the sales impact function would be to run an experiment with several treatment cells in which cost per person is exogenously varied.[25] Each treatment gives an estimate in $(c, \beta(c))$ space shown in Figure 1. A firm may use simple comparisons to measure marginal profit. Consider two spend levels $0 < c_1 < c_2$. Marginal profit is given by $m * (\beta(c_2)m - c_2) - (\beta(c_1) - c_1) = m * (\beta(c_2) - \beta(c_1)) - (c_2 - c_1)$. Estimating marginal ROI is substantially more difficult primarily because the cost differential, which can be thought of as the effective cost per exposed user, between the two campaigns $\Delta c = c_2 - c_1$ is much smaller than a standalone campaign. This means the very high standard errors given on the left side of Figure 2 are representative of the hypothesis tests required with a small Δc.[26] The difficulty in this comparison is exacerbated by the fact that the expected profit differential also decreases in Δc due to the concavity of $\beta(c)$.[27] Ideally, we would like to find c^* where this marginal profit estimate is equal to the cost of capital, but achieving such precise estimates is essentially impossible.

4 Robustness and Generalizability

In this section we discuss the robustness of our findings and look to the future about how advances in ad delivery and measurement may improve the economics of measuring the returns to advertising.

[25] See Johnson, Lewis and Reiley (2014) for an example.

[26] To see this more clearly, notice that the variance of marginal ROI has the cost differential in the denominator: $Var(ROI(\Delta c)) = \left(\frac{m}{\Delta c}\right)^2 Var\left(\beta(c_2) - \beta(c_1)\right)$.

[27] An important analog is the evaluation of ad copy. Determining if two "creatives" are significantly different will only be possible when their performance differs by a relatively wide margin.

4.1 Sales volatility

Heavily advertised categories such as high-end durable goods, subscription services such as credit cards, voting (Broockman and Green, 2013) 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. For example, for American automakers we can back out sales volatility using published data and a few back-of-the-envelope assumptions[28] conclude that the standard-deviation-to-mean-sales ratio for month-long campaign windows is 20:1, greater than that of nearly all the firms we study. In contrast, our results do not necessarily apply to new products, direct-response TV advertising or firms that get the vast majority of customers through advertising (such as a firm reliant on sponsored search). However, according to estimates from Kantar AdSpender (and other industry sources), large, publicly traded advertisers, such as the ones we study, using standard ad formats, account for the vast majority of advertising expenditure. Thus while we are careful to stress that our results do not apply to every market participant, they do have important implications for the market generally.

4.2 Scale

Our scale multipliers are designed to estimate the cost necessary to push confidence intervals to informative widths and help calibrate our findings against other, potentially more expensive advertising media (they are a lower bound when geo-randomization is the only experimentation technology available). The unfavorable economics show, however, that it would require a huge financial commitment to experimentation—the implied cost was typically in the tens of millions of dollars (and sometimes far more). Very large firms, however, often have marketing budgets that exceed these levels and, especially over time, achieving relatively precise estimates is, in principle, possible. Running repeated $500,000 experiments would allow some firms to significantly improve their understanding of the *average* impact of global spend.[29] This

[28] Purchase frequency of five years, market share of 15% and averages sales price of the 2011 median $29,793. Source: http://www.nada.org/Publications/NADADATA/2011/default.

[29] One seemingly attractive strategy is to use an evolving prior to evaluate campaigns. But as we have seen, the signal from any given campaign is relatively weak, meaning a Bayesian update would essentially return the prior. So while this is a promising strategy to determine the global average, it

type of commitment does not appear to be commonplace today, though there is at least one notable exception, Blake, Nosko and Tadelis (2013), the results of which have fundamentally shifted the advertising strategy for the firm (eBay).[30] For some large advertisers even this sort of commitment would not be enough and smaller firms may be unable to afford it.

A thought experiment, our "Super Bowl Impossibility Theorem," on advertising at scale is given the Appendix, where we consider the ability of advertisers to measure the returns for the largest reach advertising venue in the U.S. We show that even if each 30-second television commercial could be randomized at the individual level, it is nearly impossible for a firm to be large enough to afford the ad, but small enough to reliably detect meaningful differences in ROI.

4.3 Audience and expenditure intensity

The audience exposed to a firms advertisements affects not only the causal effect of the ads (the classic notion of targeting), but the precision of measurement as well. The intensity of advertising similarly impacts both quantities. For targeting, 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 by expected impact. The firm wants to design an experiment using the first M of the possible N individuals. The following derivation is straightforward so we place it in the Appendix. We find that the t-statistic on advertising impact 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.

Some firms may face hopeless trade-offs in experimenting on the entire population they wish to advertise to so instead choose to evaluate spend on a portion of individuals mostly likely to respond. Since these individuals presumably would cost more per person to advertise to, targeted tests are a natural analog to running a concentrated test in terms of higher per-person expenditure more broadly. In both cases, variance may be reduced by inducing bias in terms of extrapolating the effect to the broader user base (targeting) or for less intense expenditures (concentrated tests).

probably would not help much in evaluating any single campaign.

[30]The experiment, which utilized temporal and geographic randomization, is easily the largest to end up in published work involving tens of millions of dollars.

4.4 Advances in data and methods

Digital measurement has opened up many doors in measuring advertising effectiveness, the RCTs in this paper are certainly examples. Improving experimental infrastructure has the potential to drastically reduce the costs of running experiments. In the first generation of field experiments, major firms worked with publishers in a "high touch" fashion to implement RCTs. Advances that are already here or are on the horizon include experimentation as a service[31] and computational advertising software interfacing with real-time display and search exchanges. Both could help move the industry beyond the geographic randomization that can be currently performed "off the shelf." Ad-serving infrastructure that allows for large, free control groups (without explicit participation from the publisher) would further reduce costs and would be presumably be developed if there were sufficient demand.[32] This infrastructure could potentially incorporate pre-experiment matching as well (Deng et al., 2013).

Increasingly, more ad delivery channels are being brought into the digital fold. Early experiments with cable TV required custom infrastructure to randomize ad delivery (Lodish et al., 1995), but the ability to personalize ad delivery is reportedly being developed by major providers. Without this infrastructure, "high touch" geo-randomized advertising experiments are state of the art and experiments that rely on this method are significantly more expensive because you are unable to econometrically eliminate the noise from purchases among those who the advertiser is unable to reach with their message (similar to issues surrounding intent-to-treat). Alongside this improvement in traditional cable systems, more people are viewing TV online such as YouTube or Hulu and through devices like smart TVs, Xbox or Roku, all of which can link into digital ad serving systems. As more delivery channels fall under the experimentation umbrella, achieving the scale and justifying the financial commitment necessary to produces reliable ROI estimates becomes more realistic.

Taken together, as experiments become cheaper and easier to run and possess broader scope, the economics of measuring the returns will certainly improve. In this paper we document that the baseline these technologies will likely improve upon is

[31] Such as those offered by start-up Optimizely

[32] Technologically this requires a short serving latency between the request to the ad server, the randomization, and the request for the replacement ad. The replacement ads are known as "ghost ads"—ads that are naturally qualified to be served to a given user targeted by the campaign under study but not associated with the advertiser.

an exceedingly difficult inference problem for much of the advertising market.

5 Discussion and Conclusion

We now discuss what we believe to be the most important implications of our findings. First, since reliable feedback is scarce we expect that competitive pressure on advertising spending is weak. Consistent with this notion, Appendix Table 2 shows that otherwise similar firms (size, margins, product mix, etc.) operating in the same market often differ in their advertising expenditure by up to an order of magnitude. While this is by no means a rigorous analysis, it is consistent with the implication of our findings that very different beliefs on the efficacy of advertising can persist in the market.

The uncertainty surrounding ROI estimates can create moral hazard in communication. Suppose the "media buyer" gets a bonus based on his manager's posterior belief on campaign ROI. Applying the persuasion game model of Shin (1994), we suppose that the manager is unsure which campaigns have verifiable ROI estimates.[33] In equilibrium, the manager will be *skeptical* because she knows the media buyer will report good news when available but filter bad news, which is easy to do since experimental estimates are noisy and the truth is hard to uncover in the long-run. The manager's skepticism in turn limits the flow of information about advertising effectiveness within the firm. Up until this point of the paper, we have implicitly maintained the assumptions that 1) the firm cares about measuring ROI 2) these measurements would be reported faithfully. It turns out the inference challenge not only makes measurement difficult but can exacerbate agency problems surrounding communication.

In terms of improving the communication of results from publisher to advertiser, our experimental multipliers show that one way to reduce statistical uncertainty is to run truly massive RCTs—many 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. An increase in the demand for experimentation thus has the potential to create a new economy of

[33] Alternatively, we might suppose that estimates are always provided but the manager is unsure about which evaluation specification was used. This empirical "wiggle room" can create a similar dynamic.

scale and accordingly shape the organization of web publishing and other advertising-based industries.

Returning to our motivating question of whether the total impact of advertising justifies the aggregate expenditure in the market, our study gives a micro-founded reason as to why this is indeed an open question. A consequence is that prices and media allocations may fundamentally differ from what they would be if this question were answered because the advertising market as a whole may have incorrect beliefs about the causal impact of advertising on consumer behavior. While seemingly uncommon, potentially incorrect beliefs of this nature are not a feature unique to the advertising market. Bloom et al. (2013) argue that management consulting expenditures rarely involve a well-formed counterfactual and thus cost-effectiveness is poorly understood. In the $20 billion dollar vitamin and supplement industry, a twelve-year, 40,000-person RCT could not rule out any ex-ante reasonable impact (negative or positive) of supplements for otherwise healthy people.[34] A key difference between advertising and these industries is that advertising has a quantitative veneer the belies the true underlying uncertainty.

In conclusion, using one of the largest collections of advertising field experiments to date, we have shown that inferring the effects of advertising is exceedingly difficult. We have been careful to note that these findings do not apply to all firms or ad delivery channels, but also argued extensively that they do indeed apply to the majority of advertising dollars. We have discussed, in turn, how this informational scarcity has fundamentally shaped the advertising market. And while advances in experimentation technology will likely improve the economics of measuring advertising returns from the baseline we measure, but if realized, rather than take away from our conclusions, these technologies are likely to shape the industry's organization in ways that are directly related to the inherent measurement challenges facing firms that we have set forth.

[34]The Physicians Health Study II (Lee et al., 2005) followed 39,876 healthy women over 12 years. The 95% confidence interval on the impact of experimentally administered Vitamin E on heart attacks ranged from a 23% risk reduction to an 18% risk increase.

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.

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

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. (2013). 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.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (rev.* Lawrence Erlbaum Associates, Inc.

Damgaard, M. T. and Gravert, C. (2014). Now or never! The effect of deadlines on charitable giving: Evidence from a natural field experiment. Economics Working Papers 2014-03, School of Economics and Management, University of Aarhus.

de Bartolome, C. A. (1995). Which tax rate do people use: Average or marginal? *Journal of Public Economics*, 56(1):79–96.

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.

Gelman, A. and Carlin, J. (2013). Beyond power calculations to a broader design analysis, prospective or retrospective, using external information. *Working Paper*.

Johnson, G. A., Lewis, R. A., and Reiley, D. (2014). Location, location, location: Repetition and proximity increase advertising effectiveness. *Available at SSRN 2268215*.

Joo, M., Wilbur, K. C., Cowgill, B., and Zhu, Y. (2013). Television advertising and online search. *Management Science*, 60(1):56–73.

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

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. (In Press). Measuring the effects of advertising: The digital frontier. In Goldfarb, A., Greenstein, S., and Tucker, C., editors, *The Economics of Digitization*. NBER Press.

Lewis, R. A. and Reiley, D. H. (2014). Online advertising and offline sales: Measuring the effect of retail advertising via a controlled experiment on yahoo! *Quantitative Marketing and Economics*, page forthcoming.

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

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.

Sawyer, A. G. and Ball, A. D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research*, pages 275–290.

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

6 Appendix

6.1 Super Bowl Impossibility Theorem

We 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 expense), 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.

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:

$$R \geq \frac{C}{\gamma_C b \cdot \frac{w}{52}}. \tag{9}$$

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:

$$\mu = R \cdot \frac{\frac{w}{52}}{N_{Total}}. \tag{10}$$

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

$$R \leq \frac{N_{Total} \cdot \bar{\mu}}{\frac{w}{52}} \tag{11}$$

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:

$$\frac{C}{\gamma_C b \cdot \frac{w}{52}} \leq R \leq \frac{N_{Total} \cdot \bar{\mu}}{\frac{w}{52}}. \tag{12}$$

For the budget we choose a value, 5% of revenue, which exceeds advertising budgets for most major firms.[35] The "detectability constraint" gives the largest firm, in terms of annual revenue, that can meaningfully evaluate a given ROI hypothesis set.

Table 6.1 gives the upper and lower bounds on annual revenue. 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.[36]

Appendix Table 1: 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

[35]Source: Kantar AdSpender.

[36]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.

high and low margins respectively—but large firms or products could not. For the final two hypothesis sets, the bands are tight to vanishing. Hence the "impossibility" result.

6.2 Targeting details

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

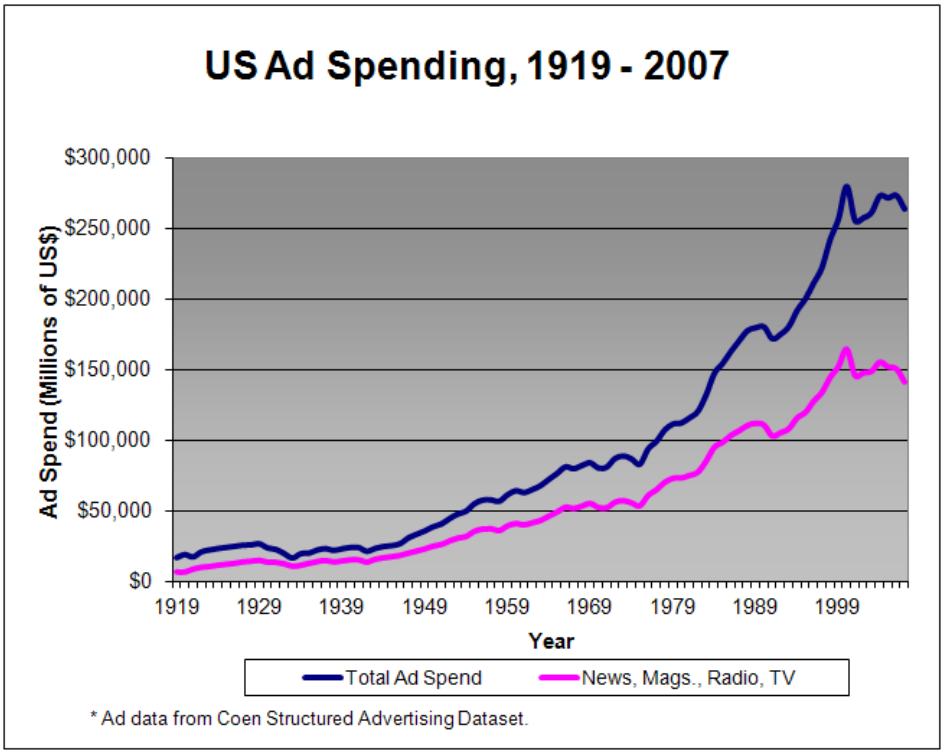
$$\begin{aligned} ROI &= \frac{\Delta\mu(M)}{C(M)} - 1 \\ \sigma^2_{ROI} &= Var\left(\frac{\Delta\mu(M)}{C(M)}\right) = \frac{2\sigma^2(M)}{M \cdot (C(M))^2} \end{aligned}$$

which implies:

$$\sigma_{ROI} = \frac{\sigma(M)}{\sqrt{M/2} \cdot C(M)} \tag{13}$$

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 US ad spending figures



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

6.4 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%