THE UNFAVORABLE ECONOMICS OF MEASURING THE RETURNS TO ADVERTISING*

RANDALL A. LEWIS AND JUSTIN M. RAO

Twenty-five large field experiments with major U.S. retailers and brokerages, most reaching millions of customers and collectively representing $2.8 million in digital advertising expenditure, reveal that measuring the returns to advertising is difficult. The median confidence interval on return on investment is over 100 percentage points wide. Detailed sales data show that relative to the per capita cost of the advertising, individual-level sales are very volatile; a coefficient of variation of 10 is common. Hence, informative advertising experiments can easily require more than 10 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 inference challenges revealed in the field experiments also show that selection bias, due to the targeted nature of advertising, is a crippling concern for widely employed observational methods. *JEL* Codes: L10, M37, C93.

I. INTRODUCTION

In the United States, firms annually spend about $500 per person on advertising (Coen 2008; Kantar Media 2012). To break even, this expenditure implies that the universe of advertisers needs to casually affect $1,500–2,200 in annual sales per person, or about $3,500–5,500 per household. A question that has remained open over the years is whether advertising affects purchasing behavior to the degree implied by prevailing advertising prices and firms' gross margins (Abraham and Lodish 1990). The rapid expansion of digital advertising—media for which ad exposure and purchasing outcomes can be measured and randomized at the individual level—has given the impression that firms will finally be able to accurately measure the returns to advertising using randomized controlled trials (RCTs). In this article, we show that although RCTs are indeed an important advance in the measurement of the returns to advertising, even

*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. We also thank countless engineers, sales people, and product managers at Yahoo!, Inc., our former employer. The work represents our views alone.

The Quarterly Journal of Economics (2015), 1941–1973. doi:10.1093/qje/qjv023.
Advance Access publication on July 6, 2015.

very large RCTs tend to produce imprecise measurements on advertising returns. The inference challenges revealed in our analysis of large-scale RCTs, in turn, exposes large biases lurking in observational methods.

Our findings are based on 25 digital advertising RCTs, accounting for $2.8 million in expenditure, from well-known retailers (19) and financial service firms (6). In each trial, or campaign, ad exposure was exogenously varied by randomly holding out a set of targeted users from receiving an advertiser's online display ad. Campaign costs ranged from $0.02 to $0.35 per exposed user, and most were close to $0.10. This corresponds to 20–60 "premium" display ads or about 7–10 prime-time television commercials.[1] Extensive data sharing agreements allow us to measure return on investment (ROI) using individual-level purchase data and gross margins provided to us by the partner firms. These data include transactions online and at brick-and-mortar stores; we can thus accurately express returns in dollar terms.[2] The detailed sales data—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 10 times the mean over typical campaign evaluation window. While this relationship may not hold true for small firms or new products, it was remarkably consistent across the relatively diverse set of advertisers in our study. Measured against this volatility, the effect on purchasing behavior required for a campaign to be profitable is very small. These two factors lead to inherently imprecise estimates of ROI: the median standard error on ROI for the retail experiments is 26.1%, implying a confidence interval over 100 percentage points wide. The median standard error for the brokerages was 115%.

These initial findings show that when advertising at a level of intensity typical of digital advertising, RCTs require sample sizes in the single-digit millions of person-weeks to distinguish campaigns that have no effect on consumer behavior (-100% ROI) from those that are profitable (ROI $>0\%$). Each experiment

1. Due to greater targeting granularity, online ads tend to have wider spreads in ad prices than other media, hence the larger range in the number of exposures for a given expenditure.

2. In particular, we sidestep "intermediate metrics," such as clicks. Later on we discuss cases in which intermediate metrics can be used with limited induced bias; however, serious complications can arise from generally using these metrics, discussed in more detail in Lewis, Rao, and Reiley (2015).

in our study had more than 500,000 unique users; most had over 1,000,000. A potential limitation, however, is that trial size was determined by the partner advertisers and may not constitute a representative sample. We overcome this drawback by computing how large each trial would have to be to reliably evaluate various hypothesis sets of interest. Specifically, we imagine an ideal scenario in which the advertiser can freely add new, independent person-weeks to their campaign.[3] The median campaign would have to be nine times larger to reliably distinguish a wildly profitable campaign (+50% ROI) from one that broke even (0% ROI). Achieving more standard tolerances for investment decisions, such as a 10% ROI difference, requires the median campaign to be 62 times larger to possess adequate power—nearly impossible for a campaign of any realistic size. Although ROI measures average returns, we briefly discuss the rather incredible difficulties in determining the average ROI target that corresponds to zero marginal profit.

Despite producing surprisingly imprecise estimates, RCTs are nonetheless a promising step forward in the science and measurement of advertising. In particular, our experimental data suggest that economically sizable biases lurk undetected in commonly employed observational methods. 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 on an indicator variable of whether 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 effect 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.[4] To successfully employ an observational

3. In practice, a firm may also opt to run an experiment with a higher "dose" of advertising. A larger per person spend makes the inference problem easier, but advertising at an undesirably high intensity can attenuate the measured ROI due to diminishing returns in the treatment effect.

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

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.[5]

In the interest of expanding the generalizability of our findings, we are careful to state our results in dollar terms: the causal sales effect based on per person advertising expenditure. This framing allows us to draw useful comparisons to heavily used media, such as television, through the lens of ad prices.[6] These comparisons allow us to credibly extend our central findings to most advertising dollars, although we are careful to discuss delivery channels, such as direct-response television commercials, and advertiser segments, such as small firms with low baseline awareness, where they are unlikely to hold.

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 further substantiate our data and methods. 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 precampaign 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

5. To see the size of selection effects, consider a simple example: if a campaign spends 10 cents per individual, which corresponds to 20–40 "premium" display ads or about 10 prime-time TV commercials, and 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%age 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.

6. In television, RCTs are becoming increasingly possible due to the digitization of distribution. For other media, geo-randomized advertising experiments are typically the state of the art. Experiments that rely on this method are significantly more expensive because they cannot eliminate the noise from purchases among those whom the advertiser is unable to reach, similar to other intent-to-treat experiments.

evaluate the applicability of our findings to specific firms or market segments.

These considerations raise a few important caveats. Our results are unlikely to apply to firms or products with low baseline sales volatility, such as those that receive nearly all their customers from advertising and products for which there is limited baseline awareness. The outlook is also more positive for campaigns that, all else equal, have substantially higher per person expenditure. However, while concentrating expenditure can increase power, diminishing returns may imply that precisely when one can measure the returns, an economically unfavorable result is more likely to occur. Media that require higher per person expenditure, such as catalogs sent via mail, are subject to an analogous problem that it may be difficult to find a suitable sample size for which the expenditure is expected to be profitable. Naturally, this drawback is less likely to apply to large firms committed to experimentation in such media. Finally, we note that ongoing advances in measurement and experimentation technology are very likely improve the inference problem advertisers face. Nonetheless, the baseline we measure indicates that imprecise beliefs on advertising effectiveness are likely to persist for most market participants in the foreseeable future.

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 despite having the access to the same technology. Second, if experimentation becomes more common, consistent with trends in the digital delivery of ads, massive publishers will be conferred an additional strategic advantage of scale, given the size of RCTs required to provide reliable feedback. Third, imprecise signals on the returns to advertising introduce issues of strategic misreporting within the firm. Finally, we note that while firms certainly make other investment decisions that have hard-to-measure returns, such as management consulting (Bloom et al. 2013) or mergers, what differs here is that the metrics and methods used in the advertising industry produce a veneer of quantitative certitude not typically found in these other circumstances. The use of RCTs to reveal the true uncertainty in measuring returns and, in the process, expose

biases nonexperimental techniques consequently has very different implications for the advertising market.

II. The Advertiser's Problem

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

II.A. 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 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 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 effect is defined by a continuous concave function of per user expenditure $\beta(c)$.[7] 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 I graphically depicts the model. c^* gives the point where the slope of the gross profit function is equal to 1 ($\beta'(c) * m = 1$). This tangency depicted is parallel to the 45-degree line, which depicts cost. c^* thus gives optimal per person expenditure. c_h give the point of intersection between the gross profit function and cost. It thus gives the spend level where ROI is exactly 0%. For any expenditure past c_h the firm has negative ROI, whereas any point to the left of c_h the firm earns positive ROI. For points in (c^*, c^h), the firm is overadvertising

7. There is evidence in support of concavity (Lewis 2010; Sahni 2013). This assumption could be weakened to "concave in the region of current spending," which essentially just says that the firm's returns to advertising are not infinite or locally convex.

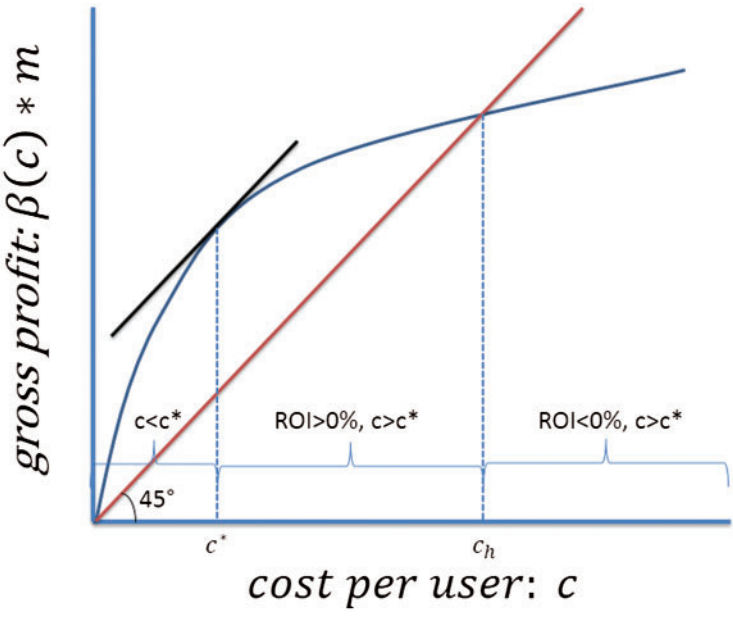

FIGURE I
The Advertiser's Problem

because marginal return is negative but the average return is still positive. For points below c^* marginal profit exceeds marginal cost, meaning the firm is underadvertising, but again ROI is positive.

The model formalizes the estimation of the average per person effect 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 experience in the industry.

II.B. 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."

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 to 23, but tends to be clustered between 10 and 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.[8]

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 effect 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

8. In contrast, homogeneous foodstuffs 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).

conditioning on exogenous and predetermined factors, such as precampaign sales, that are predictive of baseline purchases or that may have differed by chance between treatment and control subjects. We suppress the additional covariates later 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 on" caveat where the R^2 becomes partial R^2 of the treatment variable.

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 effect estimate, β, can be converted to ROI by multiplying by the gross margin to get the gross profit effect, 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.

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

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

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

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

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. In the model with

covariates, the partial R^2 after first conditioning on covariate (see Lovell 2008 for a thorough explanation of this algebra):

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

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 for testing the hypothesis ($\beta = 0$) as a function of this ratio as well:

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

We call $(\frac{\Delta\bar{y}}{\sigma})$ the impact-to-standard-deviation ratio.[9]

We calibrate these statistics using a representative experiment—slightly larger than the median—from our study. For ease of exposition, we discuss the hypothetical case as if it were a single, actual experiment. The cost per exposed user is \$0.14, which corresponds to roughly 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 of \$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.[10] From our derivation above, this implies an R^2 of:

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

9. It is also known as Cohen's d.

10. This is less than $\frac{1}{40}$ the "small" effect size of 0.2 outlined in Cohen (1977).

Perhaps surprisingly, even a very successful campaign has a minuscule R^2 of 0.0000054.[11] An immediate consequence is that a very large N is required to reliably distinguish the effect 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, or zero causal effect, 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 alternative hypothesis of a healthy 25% ROI, we fail to reject the null that the ad had no causal effect 74% of the time.[12]

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 (Reiley, Rao, and Lewis 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 a few percentage points higher baseline purchase rate can lead to an

11. This is less than $\frac{1}{10,000}$ 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.

12. 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).

expected bias many multiples of the treatment effect. Unless this difference is controlled for with near perfect precision, observational models will have large biases.

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 coauthored by the president of comScore, one of the largest data providers for web publishers and advertisers, reported a 300% effect 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, Nosko, and Tadelis (2014).

III. The 25 Randomized Controlled Trials

In this section we delve into RCTs. There is an inherent challenge in discussing this many experiments in adequate detail. Our strategy is to put detailed information in three comprehensive summary tables and focus our discussion in the text to aggregated measures, highlighting underlying heterogeneity where appropriate. Finally, we limit our reporting to the statistical uncertainty surrounding the measurement of advertising returns and do not report the point estimates for each campaign. We do so for a few reasons. First, although 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, raising the possibility of bias (or lack of generalizability) for any findings thereto. 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.

III.A. Overview and Data Description

Table I gives key summary statistics for the 25 display advertising experiments. All experiments took place between

TABLE I

SUMMARY OF THE 25 ADVERTISING FIELD EXPERIMENTS

Panel A: Retailers: In-store + online sales

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Campaign summary						Per customer		
			Assignment[a]		Exposed				
Adv. expt.	Days	Cost ($K)	Test	Control	Test	Control	μ sales ($, control)	σ sales ($)	$\frac{\sigma}{\mu}$
1.1	14	129	1,257,756	300,000	814,052	—	9.49	94.28	9.9
1.2	10	40	1,257,756	300,000	686,878	—	10.50	111.15	10.6
1.3	10	68	1,257,756	300,000	801,174	—	4.86	69.98	14.4
1.4	105	260	957,706	300,000	764,235	238,904	125.74	490.28	9.7
1.5	7	81	2,535,491	300,000	1,159,100	—	11.47	111.37	3.9
1.6	14	150	2,175,855	1,087,924	1,212,042	604,789	17.62	132.15	7.5
2.1	35	192	3,145,790	3,146,420	2,229,959	—	30.77	147.37	4.8
2.2	35	19	3,146,347	3,146,420	2,258,672	—	30.77	147.37	4.8
2.3	35	19	3,145,996	3,146,420	2,245,196	—	30.77	147.37	4.8
3.1	3	10	281,802	161,163	281,802	161,163	1.27	18.46	14.6
3.2	4	17	483,015	277,751	424,380	—	1.08	14.73	13.7
3.3	2	26	292,459	169,024	292,459	169,024	1.89	18.89	10.0
3.4	3	18	311,566	179,709	311,566	179,709	1.29	16.27	12.6
3.5	3	18	259,903	452,983	259,903	—	1.75	18.60	10.6
3.6	4	27	355,474	204,034	355,474	204,034	2.64	21.60	8.2
3.7	2	34	314,318	182,223	314,318	182,223	0.59	9.77	16.6
4.1	18	90	1,075,828	1,075,827	693,459	—	0.56	12.65	22.6
5.1	41	180	2,321,606	244,432	1,583,991	—	54.77	170.41	3.1
5.2	32	180	600,058	3,555,971	457,968	—	8.48	$70.20	8.3
Mean	19.8	100	1,325,078	975,279	902,454	—	18.23	$95.94	10.0
Median	10.0	81	1,075,828	300,000	693,459	—	8.48	$70.20	9.9

Panel B: Financial services: New account sign-ups, online only

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Campaign summary						Per customer		
			Assignment[b]						
Adv. expt.	Days	Cost ($K)	Test	Control	Exposed	New accts	μ New (test)	σ New acct.	$\frac{\sigma}{\mu}$
6.1	42	50	12%	52%	794,332	867	0.0011	0.0330	30.3
6.2	42	50	12%	52%	748,730	762	0.0010	0.0319	31.3
6.3	42	75	12%	52%	1,080,250	1,254	0.0012	0.0341	29.3
6.4	42	75	12%	52%	1,101,638	1,304	0.0012	0.0344	29.0
7.1	42	613	90%	10%	17,943,572	10,263	0.0006	0.0239	41.8
7.2	36	86	8,125,910	8,125,909	793,042	1090	0.0014	0.0331	24.1
Mean	41.0	158	—	—	3,743,594	2,590	0.0011	0.0317	31.0
Median	42.0	75	—	—	937,291	1,172	0.0011	0.0331	29.8

[a]Counts represent the number of unique user identifiers either assigned to the experiment or exposed by the advertiser's campaign.

[b]"X%" defines the fraction of all Yahoo! users who were eligible for participation in an experiment that randomized "on the fly."

2007 and 2011. We are not permitted to disclose the identity of the advertisers. They are large retailers (Panel A) and financial service firms (Panel B) that are most likely familiar to North American readers. We employ a naming convention of *Advertiser*. *Experiment* as identifiers; these are given in column (1).[13] Means and medians for all relevant values are given at the bottom of each panel.

Column (2) gives the campaign duration, which ranged from 2 to 105 days with a median of 10 days for retailers, 42 days for financial service firms, and 14 days for the pooled data. These durations are consistent with standard industry practice. Column (3) gives the cost of the campaign, the median $85,000 for retailers and $75,000 for financial firms, while the range varied from relatively small ($9,964) to quite large ($612,693). Combined, the campaigns represent over $2.8 million in expenditure. Later on we convert overall expenditure to cost per exposed user, which is a more meaningful measure because it can be easily compared to the cost of other advertising media and conveys the advertising "dose."

The median campaign reached over 1 million individuals, and all campaigns had hundreds of thousands of individuals in both test and control cells, as shown in columns (4)–(7) (both panels). The distinction between "assignment" and "exposed" is due to the fact that all the users assigned to a treatment or control group did not actually arrive at the web publisher to be served an ad. In some experiments, control users were served an ad for a charity, whereas in others they were simply held out of the treatment group (and hence the exact size of the "treated control" is not known, see Lewis and Schreiner 2010 for a more complete discussion of this aspect of experimental design).

Columns (8)–(10) give revenue data at the customer level. In the top panel, revenue comes in the form of the sale of physical goods and is thus denoted in dollars. In the bottom panel, revenue comes by monetizing customers who sign up for an online trading account. In the interest of simplicity, we discuss these cases separately. Column (8) gives sales per customer in the control group for the campaign evaluation period. Median sales was $8.48 (mean $18.23). Sales per customer varies across retailers and

13. Many of the experiments are taken from past work from Yahoo! Labs, like Lewis and Reiley (2014); Reiley, Lewis, and Schreiner (2012); Johnson, Lewis, and Reiley (2015); and Reiley, Rao, and Lewis (2011).

campaigns, which is due to differing firm popularity, campaign length, and the degree of targeting used in the campaign (a more targeted campaign typically has higher baseline sales). Column (9) gives the standard deviation of sales per customer. The median is \$70.20 (mean \$95.54). Column (10) gives the standard deviation to mean ratio, which we have seen is an important quantity in our analysis. The median ratio is 9.9 and it exceeds 7 for all but two experiments. Longer campaigns tend to have slightly lower 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 preperiod sales, which is a strategy we employ in estimation.[14]

For financial service firms, column (8) gives the mean probability of an account sign-up. The median is 0.0011, reflecting a low base rate, even for users targeted by the campaign, for signing up for an online brokerage account. Column (9) gives the standard deviation of sign-up rate, and column (10) gives the more interpretable standard deviation to mean ratio. The median ratio is 29.8 and all exceed 24, which is larger than every retailer campaign. These high values reflect the all-or-nothing nature of acquiring a long-term customer.

III.B. Estimating the Returns to Advertising

The first step in measuring the returns to advertising is defining evaluation windows for the statistical analysis. In working with our partner firms, we followed standard industry practice of including the period of time ads were running and a relatively short window, 1–4 weeks, following the campaign. In principle, however, the effects of advertising could be very long-lived and therefore the bias-minimizing window would be correspondingly long. Although it may seem counterintuitive at first, long windows tend to damage statistical power. In Lewis, Rao, and Reiley (2015), we establish the following condition: if the next week's expected causal effect is less than one-half the average casual effect over all previous weeks, then including it reduces

14. 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. Furthermore, while longer campaigns generate more points of observation, these additional data will only make inference easier as long as the spending per person per week is not diluted.

the t-statistic of the total treatment effect. The proposition tells us when a marginal week hurts estimation precision by contributing more noise than signal. Unless there is limited decay in the ad effect over time, short windows are optimal from a power perspective.[15] The campaign windows we choose should thus be viewed as erring in favor of variance reduction at the possible expense of bias. For "buy it now or never" response patterns (Damgaard and Gravert 2014), this bias will be minimal, otherwise our approach will overestimate statistical precision to some degree.

With evaluation windows in hand, we are now ready to estimate the causal effects of advertising. Table II summarizes the statistical inference for each campaign. We start with a more detailed description of the data and protocol. Column (2) gives the unit of observation for sales and brokerage account sign-ups. For the retailers, 10 experiments have daily aggregation (indicated by 1), five are weekly (2), and four encompass the entire campaign period (3). For financial firms, account sign-ups were aggregated for the campaign period except in one case, where it was observed daily. Column (3) gives details of how the experiments were implemented by the web publisher. Every experiment had a randomized set of users for the control group designated by a 1. Those experiments (the majority) that ensured the control group was currently active on the web publisher are designated with a 2; experiments where placebo ads were explicitly shown, as opposed to simply ensuring the firm's ad was not shown, are designated by a 3; and experiments designated with a 4 had multiple treatments. Column (4) indicates how the sales data were filtered to improve efficiency.

Column (5) reports the control variables we have available to improve the precision of our experimental estimates. These include lagged sales (indicated by 1), demographics (2), and online behaviors (such as intensity of browsing, 3), all measured at the individual level. Lagged sales were available for 16 retail experiments. Lagged sales are not available for brokerages since these campaigns were specifically designed to get users to open new trading accounts. Demographics were available for six

15. As an example, suppose the causal effect of the advertising on weeks 1, 2, and 3 is 5%, 2%, and z%, respectively. Then z must be greater than $\frac{5+2}{2}/2 = 1.75$. 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.

experiments, and online behavior was available for all experiments except those run by firm 2.

To estimate the returns to advertising for each experiment, we used the appropriate panel techniques to predict and absorb residual variation using the data given in column (5). We find that lagged sales are the best predictor, reducing variance in the dependent variable by as much as 40%. The magnitude of these reductions and the fact that lagged sales is the best predictor are consistent with related work (Deng et al. 2013). A more useful frame for these reductions is their impact on the standard error on the treatment variable. 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%.[16] Indeed, to achieve an order-of-magnitude reduction in standard errors, one would have to predict sales with an $R^2_{|\mathbf{W}} = 0.99$.

The second group of columns in Table II gives the key statistical properties for each campaign. Column (6) 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 in order to obtain as precise an estimate as possible. In column (7), we translate this into the implied radius (± window) of the 95% confidence interval for the sales effect, in percentage terms of baseline sales. The median radius is 5.5%. Column (8) gives the per person advertising spend, which ranges from \$0.02 to \$0.39 and is centered on \$0.10. This figure provides a useful benchmark not only for online advertising expenditures but also for media purchases in other advertising channels.

In column (10) we translate the sales effect standard errors to ROI using our estimates of gross margins in column (9) for retailers and lifetime customer values for financial firms. Using a fixed lifetime customer value is not ideal because it restricts advertising from affecting potentially heterogeneous value extracted over the life span of an acquired customer. More generally, the assumption of advertising acting through a single channel is 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

16. $1 - \sqrt{\frac{1-R^2_{\mathbf{W}}}{1-R^2}} = 1 - \sqrt{1 - R^2_{|\mathbf{W}}} = 23\%$.

TABLE II

ESTIMATION FOR THE 25 ADVERTISING FIELD EXPERIMENTS

Panel A: Retailers: In-store + online sales

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Estimation strategies				Key campaign statistics				
Adv. expt.	Y[a]	X[b]	Y&X[c]	W[d]	Std. err. β sales ($)	1.96*std. err. % sales	Spend/ exposed ($)	Gross margin (%)	Std. err. ROI (%)
1.1	2	1	—	1,2,3	$0.193	4.0	0.16	50	61
1.2	2	1	—	1,2,3	$0.226	4.2	0.06	50	193
1.3	2	1	—	1,2,3	$0.143	5.8	0.09	50	84
1.4	2	1,2,3	—	1,2,3	$0.912	1.4	0.34	50	134
1.5	2	1,2	—	1,2,3	$0.244	4.2	0.04	50	278
1.6	1	1,2,3,4	1	1,2,3	$0.207	2.3	0.12	50	84
2.1	3	1	—	—	$0.139	0.9	0.09	15	24
2.2	3	1	—	—	$0.142	0.9	0.08	15	25
2.3	3	1	—	—	$0.131	0.8	0.09	15	23
3.1	1	1,2,3	1	1,3	$0.061	9.5	0.04	30	52
3.2	1	1,2	1	1,3	$0.044	8.0	0.04	30	34
3.3	1	1,2,3	1	1,3	$0.065	6.7	0.09	30	22
3.4	1	1,2,3	1	1,3	$0.051	7.8	0.06	30	26
3.5	1	1,2	1	1,3	$0.049	5.5	0.07	30	21
3.6	1	1,2,3	1	1,3	$0.064	4.8	0.08	30	25
3.7	1	1,2,3	1	1,3	$0.032	10.6	0.11	30	9
4.1	1	1,2	1	1	$0.031	10.9	0.13	40	10
5.1	3	1,2	—	1,3	$0.215	0.8	0.11	30	57
5.2	1	1,2	1	1,3	$0.190	4.4	0.39	30	15
Mean	—	—	—	—	$0.165	4.9	0.11	34	62
Median	—	—	—	—	$0.139	4.4	0.09	30	26

Panel B: Financial services: New account sign-ups, online only

Adv. expt.	Y[a]	X[b]	Y&X[c]	W[d]	Std. err. β[e] new accts	1.96*std. err. % new	Spend/ person ($)	Lifetime value ($)	Std. err. ROI (%)
6.1	3	1,2,4	—	3	69	15.6	0.06	1,000	138
6.2	3	1,2,4	—	3	69	17.7	0.07	1,000	137
6.3	3	1,2,4	—	3	70	10.9	0.07	1,000	93
6.4	3	1,2,4	—	3	70	10.5%	0.07	1,000	93
7.1	1	1,2,3,4	1,2	3	288	5.5	0.03	1,000	47
7.2	3	1,2	—	3	46	8.3	0.02	1,000	233
Mean	—	—	—	—	102	11.4	0.05	1,000	123
Median	—	—	—	—	69	10.7	0.06	1,000	115

[a]Y: 1: daily, 2: weekly, 3: total campaign window.
[b]X: 1: randomized control, 2: active on Yahoo! or site where ads were shown, 3: placebo ads for control group, 4: multiple treatments.
[c]Y&X: 1: sales filtered post first exposure or first page view, 2: outcome filtered based on postexposure time window.
[d]W: 1: lagged sales, 2: demographics, 3: online behaviors.
[e]β here is the causal effect rescaled in terms of new accounts—approximately the regression coefficient on ad exposure times the number of exposed users. Std. err. denotes standard errors; ROI denotes return on advertising investment.

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, Lewis, and Reiley 2015), 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 response 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 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.

Returning to column (10), 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*. Given all but the least informative priors, such an interval is 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 nature of long-term customer acquisition) and accordingly had higher standard errors on ROI, with all but one campaign exceeding 93%.

In Figure II we plot the standard error of the ROI estimate against per capita campaign cost. Each line represents a different advertiser. Two important features are immediately apparent. First, there is significant heterogeneity across firms. Firm 1 and the financial firms have the highest statistical uncertainty in the ROI estimate. We have already discussed why this is the case for financial services; firm 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, high expenditure per person induces a large dose of advertising. This in turn means that a larger impact on sales must occur to deliver

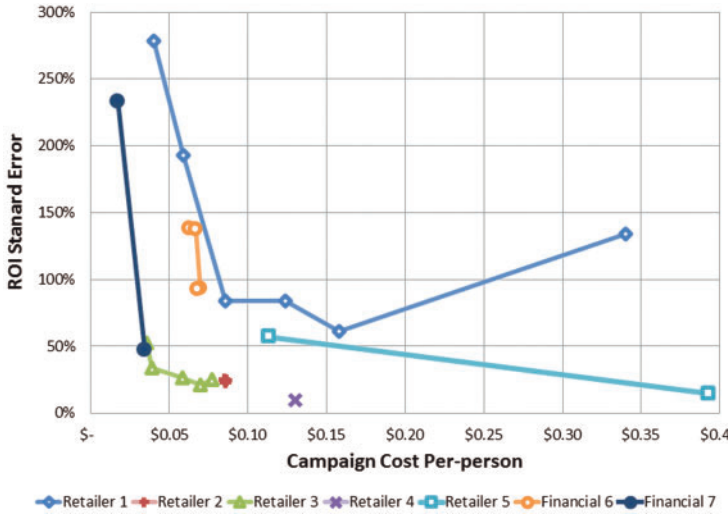

FIGURE II

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 larger the dose, the better the power. The most intense spend was firm 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 preperiod sales and online behaviors to condition on. Despite these favorable design elements, the 95% confidence interval on ROI is still 60 percentage points wide.

Figure II 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. A firm can improve precision by making a bias-variance trade-off that involves intentionally using more intense expenditure than it would in normal practice. Conversely, for firms for which low spending per person is likely to be optimal, then measuring returns will be substantially more difficult. For instance, the figure shows that RCTs with an expenditure below $0.05 per person offered uninformative estimates for these advertisers. In the next section we will see that most of these low-dose experiments would have needed to be more than 10 times larger to

reliably evaluate a null hypothesis of zero effect. In section III.D, we discuss how the difficulties in evaluating low expenditure campaigns implies an analogous difficulty in distinguishing campaigns of similar expenditure per person, which is a necessary step in optimization.

III.C. Statistical Power and Experiment Counterfactuals

In Table III we compute how much larger the experiments would have to be to reliably evaluate various sets of hypotheses on the returns to advertising investments. The multipliers get to the heart of the economics of measuring the returns to advertising by defining 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 disparate null and alternative hypotheses and then draw the hypotheses closer to tolerances more typical of investment decisions. For each hypothesis set, we first give the expected t-statistic, $E[t]$, to reject the null hypothesis when the true state of the world is given by the alternative hypothesis. This reflects the expected statistical significance of the actual experiment provided the alternative is true. The "experiment multiplier" tells us how much larger an experiment would have to be in terms of new, independent individuals to achieve adequate power, which we define as an expected t-statistic of 3, as this produces power of 91% with a one-sided test size of 5%. 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.[17]

17. 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. Finally, we note that running more concentrated tests can also improve power, an issue we discuss in section IV.C.

TABLE III

STATISTICAL POWER OF THE 25 ADVERTISING EXPERIMENTS

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
		Ads did anything? H_0: ROI = −100% H_a: ROI = 0% ΔROI = 100%		Highly profitable? H_0: ROI = 0% H_a: ROI = 50% ΔROI = 50%		Strong performer? H_0: ROI = 0% H_a: ROI = 10% ΔROI = 10%		Maximized profits? H_0: ROI = 0% H_a: ROI = 5% ΔROI = 5%	
Adv. expt.	Std. err. ROI (%)	$E[t]$[a]	$N\times$ for $E[t]$=3[b] (×)	$E[t]$	$N\times$ for $E[t]$=3 (×)	$E[t]$	$N\times$ for $E[t]$=3 (×)	$E[t]$	$N\times$ for $E[t]$=3 (×)
Panel A: Retailers: In-store + online sales									
1.1	61	1.64	3.3	0.82	13.4	0.16	335	0.08	1,338
1.2	193	0.52	33.5	0.26	133.8	0.05	3,345	0.03	13,382
1.3	84	1.19	6.3	0.60	25.2	0.12	631	0.06	2,524
1.4	134	0.75	16.2	0.37	64.7	0.07	6,939	0.02	27,756
1.5	278	0.36	69.4	0.37	277.6	0.07	425	0.07	1,700
1.6	84	1.20	6.3	0.60	25.2	0.12	629	0.06	2,515
2.1	24	4.12	0.5	2.06	2.1	0.41	53	0.21	212
2.2	25	3.99	0.6	2.00	2.3	0.40	57	0.20	226
2.3	23	4.33	0.5	2.17	1.9	0.43	48	0.22	192
3.1	52	1.92	2.4	0.96	9.7	0.19	243	0.10	972
3.2	34	2.96	1.0	1.48	4.1	0.30	103	0.15	411
3.3	22	4.50	0.4	2.25	1.8	0.45	44	0.23	177
3.4	26	3.82	0.6	1.91	2.5	0.38	62	0.19	247
3.5	21	4.73	0.4	2.36	1.6	0.47	40	0.24	161
3.6	25	3.98	0.6	1.99	2.3	0.40	57	0.20	227
3.7	9	11.32	0.1	5.66	0.3	1.13	7	0.57	28
4.1	10	10.45	0.1	5.22	0.3	1.04	8	0.52	33
5.1	57	1.76	2.9	0.88	11.6	0.18	291	0.09	1,165
5.2	15	6.90	0.2	3.45	0.8	0.69	19	0.34	76
Mean	62	3.71	7.6	1.86	30.6	0.37	765	0.19	3,058
Median	26	3.82	0.6	1.91	2.5	0.38	62	0.19	247
Panel B: Financial services: New account sign-ups, online onlys									
6.1	138	0.73	17.1	0.36	68.3	0.07	1,707	0.04	6,828
6.2	137	0.73	17.0	0.36	67.9	0.07	1,697	0.04	6,790
6.3	93	1.07	7.8	0.54	31.4	0.11	785	0.05	3,139
6.4	93	1.08	7.7	0.54	30.9	0.11	77	0.05	3,094
7.1	47	2.13	2.0	1.06	8.0	0.21	199	0.11	795
7.2	233	0.43	48.7	0.21	195.0	0.04	4,874	0.02	19,496
Mean	123	1.03	16.7	0.51	66.9	0.10	1,673	0.05	6,690
Median	115	0.90	12.4	0.45	49.6	0.09	1,241	0.04	4,964

[a] $E[t]$ is the expected t-statistic for testing the null versus the alternative hypothesis, given the ROI standard error.

[b] Equal to $(\frac{3}{E[t]})^2$—the multiple of the experiment's sample size, N, necessary for a powerful experiment where $E[t] = 3$.

In columns (3)–(4) 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, 2007).[18] Nine experiments had $E[t] < 1.65$ (column (3)), meaning the most likely outcome was failing to reject -100% ROI when the truth was the ad was profitable.[19] The multipliers indicate that these experiments would have to be substantially larger to produce informative results. 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 underpowered and could reliably evaluate this hypothesis set with a modest increase in size.

Simply rejecting that a campaign was a total waste of money is not an ambitious goal. In columns (5)–(6) 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 12 experiments had $E[t] < 1$ (severely underpowered), 4 had $E[t] \in [1, 2]$, 5 had $E[t] \in [2, 3]$ ($90\% >$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 seven could reach this mark by increasing the size of the experiment by a factor of about 2.5 (those with $E[t] \in [1.9, 3]$) 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 firm 5's second campaign, which cost \$180,000 and reached 457,968 people, and firm 4's campaign, which cost \$90,000 and reached 1,075,828 people. For firm 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 II. Firm 4 had good power due to two key

18. 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 and Rao 1989), and online ads (Reiley, Rao, and Lewis 2011; Lewis and Reiley 2014).

19. 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.

factors: it had the fourth highest per person spend and the second lowest standard deviation of sales.

Identifying highly successful campaigns from ones that merely 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. In columns (7)–(8) we draw the hypotheses to a more standard tolerance of 10 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. 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 seven times larger to have sufficient power to distinguish this difference. The median retail sales experiment would have to be 61 times larger (with nine exceeding 100×) 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 1,241.

In columns (9)–(10) we push the envelope further, setting the difference between the test hypotheses to 5 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.[20]

Many investment decisions involve underlying uncertainty. In drug discovery, for example, a handful of drugs like Lipitor (atorvastatin, Pfizer) 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

20. We are trying to estimate $\frac{1}{26}$th of the previous effect size, which is 26^2 times harder.

this very reason, and ex post profit measurement is relatively straightforward. Advertisers tend to vary ad copy and campaign style to diversify expenditure, and although 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 it is not possible to make precise ROI statements. 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 substantially better than the levels of efficiency we report (conditional on sample size, etc.) should be viewed with skepticism.

III.D. Optimizing Expenditure with a ROI Target

Returning to Figure I, 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 underadvertising—ROI is too high. For $c > c^*$, the firm is overadvertising: 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.[21] Each treatment gives an estimate in $(c, \beta(c))$ space shown in Figure I. A firm may use simple comparisons to measure

21. See Johnson, Lewis, and Reiley (2015) for an example.

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 II are representative of the hypothesis tests required with a small Δc.[22] The difficulty in this comparison is exacerbated by the fact that the expected profit differential also decreases in Δc due to the concavity of $\beta(c)$.[23] 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.

IV. Robustness and Generalizability

In this section, we discuss the robustness of our findings and look to the future on how technological advances have the potential to improve the economics of measuring the returns to advertising.

IV.A. Sales Volatility

Heavily advertised categories such as high-end durable goods, subscription services such as credit cards, voting (Broockman and Green 2014), 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 U.S. automakers we can back out sales volatility using published data and a few back-of-the-envelope assumptions[24] conclude that the standard-deviation-to-mean-sales ratio

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

23. 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.

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

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.

IV.B. 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 possible, in principle. Running repeated $500,000 experiments would allow some firms to significantly improve their understanding of the average impact of global spend.[25] This type of commitment does not appear to be commonplace today, though there is at least one notable exception: Blake, Nosko, and Tadelis (2014), the results of which shifted the advertising strategy for the firm (eBay).[26] 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 in the Online Appendix, where we consider the ability of advertisers to measure

25. 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 probably would not help much in evaluating any single campaign.

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

the returns for the largest reach advertising venue in the United States. 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 still small enough to reliably detect meaningful differences in ROI.

IV.C. Audience and Expenditure Intensity

The audience exposed to a firm's 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 affects 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 individuals 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 Online 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 empirical.

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).

IV.D. Advances in Data Collection and Methods

Digital measurement has opened up many doors in measuring advertising effectiveness; the RCTs in this article are examples. Improving experimental infrastructure has the potential to dramatically 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 have recently arrived include experimentation as a service 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 presumably would be developed if there were sufficient demand. 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 television 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 they cannot econometrically eliminate the noise from purchases among those whom the advertiser is unable to reach with their message (as intent-to-treat estimators). 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 article we document that the baseline these technologies will likely improve on is a difficult inference problem for much of the advertising market. The historical difficulty of this problem means that emerging experimentation technologies have the potential to disrupt the market despite the challenges we document in running informative RCTs.

V. 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, Table II in the Online Appendix 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. Although 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.[27] 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. These considerations highlight that up until this point of the article we have implicitly maintained two important assumptions: (i) the firm cares about measuring ROI, and (ii) these measurements would be reported faithfully. Models of strategic communication show that inference challenges not only makes measurement itself difficult but can exacerbate agency problems.

In terms of improving the precision of results a publisher can provide to advertisers, our experimental multipliers show that one way to reduce statistical uncertainty is to run massive RCTs. These multipliers indicate advertisers could narrow confidence intervals to an acceptable tolerance with well-designed experiments in the tens of millions of users in each treatment cell. Larger advertisers committed to experimentation could afford such RCTs, but only the largest publishers could provide the scale required. An increase in the demand for experimentation has the potential to create a new economy of scale and accordingly shape the organization of advertising-based industries.

Returning to the motivating question we posed at the outset, our study gives a micro-founded reason as to why it is indeed unclear whether the total impact of advertising justifies the aggregate expenditure in the market. A consequence is that prices and media allocations may substantially differ from what they would be with more complete information. Put simply, the

27. Alternatively, we might suppose that estimates are always provided but the manager is unsure about which evaluation specification was used. This type of methodological "wiggle room" can create a similar dynamic.

advertising market as a whole may have incorrect beliefs about the causal impact of advertising on consumer behavior. Increasing adoption of experimentation and cost reductions due to technological advances thus have a disruptive potential for this market. Potentially incorrect beliefs of this nature are not a feature unique to the advertising market, although they do appear to be uncommon in well-functioning market economies. Bloom et al. (2013) argue that management consulting expenditures rarely involve a well-formed counterfactual and thus real returns are poorly understood. In the $20 billion vitamin and supplement industry, a 12-year, 40,000-person RCT could not rule out any ex ante reasonable impact (negative or positive) of supplements for otherwise healthy people.[28] A key difference between advertising and these industries is that advertising has a quantitative veneer that belies the true underlying uncertainty.

In conclusion, using one of the largest collections of advertising RCTs 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 we have argued that they do apply to the majority of advertising dollars. We have discussed how this informational scarcity has fundamentally shaped the advertising market. Advances in experimentation technology will likely improve the economics of measuring advertising returns from the baseline we measure, but if realized these technologies, rather than take away from our conclusions, 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.

GOOGLE
MICROSOFT RESEARCH

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at QJE online (qje.oxfordjournal.org).

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

REFERENCES

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

Abraham, Magid, and Leonard Lodish, "Getting the Most out of Advertising and Promotion," *Harvard Business Review*, 68 (1990), 50.

Bagwell, Kyle, "The Economic Analysis of Advertising," in *Handbook of Industrial Organization,* Vol. 3, Mark Armstrong and Robert H. Porter, eds. (Amsterdam: North-Holland, 2007).

Blake, Tom, Chris Nosko, and Steven Tadelis, "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment," NBER Working Paper w20171, 2014

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts, "Does Management Matter? Evidence from India," *Quarterly Journal of Economics*, 128 (2013), 1–51.

Broockman, David, and Donald Green, "Do Online Advertisements Increase Political Candidates' Name Recognition or Favorability? Evidence from Randomized Field Experiments," *Political Behavior*, 36 (2014), 263–289.

Card, David, "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, vol. 3, Orley Ashenfelter and David Card, eds. (Amsterdam: Elsevier, 1999).

Carroll, Vincent, Ambar Rao, Hau Lee, Arthur Shapiro, and Barry Bayus, "The Navy Enlistment Marketing Experiment," *Marketing Science*, 4 (1985), 352–374.

Coen, Richard, *Coen Structured Advertising Expenditure Dataset*, 2008, available at http://purplemotes.net/2008/09/14/us-advertising-expenditure-data/.

Cohen, Jacob, *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, NJ: Lawrence Erlbaum, 1977).

Damgaard, Mette, and Christina Gravert, "Now or Never! The Effect of Deadlines on Charitable Giving: Evidence from a Natural Field Experiment," Economics Working Paper, School of Economics and Management, University of Aarhus, 2014.

de Bartolome, Charles, "Which Tax Rate Do People Use: Average or Marginal?," *Journal of Public Economics*, 56 (1995), 79–96.

Deng, Alex, Ya Xu, Ronny Kohavi, and Toby Walker, "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*, 2013.

Eastlack, Joseph, Jr., and Ambar Rao, "Advertising Experiments at the Campbell Soup Company," *Marketing Science*, 8 (1989), 57–71.

Gelman, Andrew, and John Carlin, "Beyond Power Calculations to a Broader Design Analysis, Prospective or Retrospective, Using External Information," Columbia University Working Paper, 2013.

Johnson, Garrett, Randall A. Lewis, and David H. Reiley, "Location, Location, Location: Repetition and Proximity Increase Advertising Effectiveness," SSRN Working Paper 2268215, 2015.

Joo, Mingyu, Kenneth Wilbur, Bo Cowgill, and Yi Zhu, "Television Advertising and Online Search," *Management Science*, 60 (2013), 56–73.

Kaiser, Harry, *Economics of Commodity Promotion Programs: Lessons from California* (New York: Peter Lang, 2005).

Kantar Media, Kantar AdSpender, 2012, available at http://www.kantarmedia.us/product/adspender.

Lee, I-Min, Nancy Cook, J. Michael Gaziano, David Gordon, Paul Ridker, JoAnn Manson, Charles Hennekens, and Julie Buring, "Vitamin E in the Primary Prevention of Cardiovascular Disease and Cancer," *Journal of the American Medical Association*, 294 (2005), 56.

Lewis, Randall A., *Where's the "Wear-Out?": Online Display Ads and the Impact of Frequency*, PhD diss., Massachusetts Institute of Technology, 2010.

Lewis, Randall A., Justin M. Rao, and David H. Reiley, "'Measuring the Effects of Advertising: The Digital Frontier," in *Economic Analysis of the Digital Economy*, Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, eds. (Chicago: NBER Press, 2015).

Lewis, Randall A. and Taylor A. Schreiner, "Can Online Display Advertising Attract New Customers?," *MIT Ph.D. Thesis*, 2010.
Lewis, Randall A., and David H. Reiley, "Online Ads and Offline Sales: Measuring the Effects of Retail Advertising via a Controlled Experiment on Yahoo!," *Quantitative Marketing and Economics*, 12 (2014), 235–266.
Lodish, Leonard, Magid Abraham, Stuart Kalmenson, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens, "How TV Advertising Works: A Meta-Analysis of 389 Real World Split Cable TV Advertising Experiments," *Journal of Marketing Research*, 32 (1995), 125–139.
Lovell, Michael, "A Simple Proof of the FWL Theorem," *Journal of Economic Education*, 39 (2008), 88–91.
Reiley, David H., Randall A. Lewis, and Taylor Schreiner, "Ad Attributes and Attribution: Large-Scale Field Experiments Measure Online Customer Acquisition," unpublished manuscript, 2012.
Reiley, David H., Justin M. Rao, and Randall A. Lewis, "Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising," in *Proceedings of the 20th ACM International World Wide Web Conference*, 2011.
Sahni, Navdeep, "Effect of Temporal Spacing between Advertising Exposures: Evidence from Online Field Experiments," Stanford GSB Working Paper, 2013.
Sawyer, Alan, and A. Dwayne Ball, "Statistical Power and Effect Size in Marketing Research," *Journal of Marketing Research*, 18 (1981), 275–290.
Shin, H. S., "News Management and the Value of Firms," *RAND Journal of Economics*, 25 (1994), 58–71.

This page intentionally left blank