

Improving the Effectiveness of Time-Based Display Advertising

DANIEL G. GOLDSTEIN and R. PRESTON MCAFEE and SIDDHARTH SURI,

Yahoo! Research

Display advertisements are typically sold by the impression, where one impression is simply one download of an ad. Previous work has shown that the longer an ad is in view, the more likely a user is to remember it and that there are diminishing returns to increased exposure time [Goldstein et al. 2011]. Since a pricing scheme that is at least partially based on time is more exact than one based solely on impressions, time-based advertising may become an industry standard. We answer an open question concerning time-based pricing schemes: how should time slots for advertisements be divided? We provide evidence that ads can be scheduled in a way that leads to greater total recollection, which advertisers value, and increased revenue, which publishers value. We document two main findings. First, we show that displaying two shorter ads results in more total recollection than displaying one longer ad of twice the duration. Second, we show that this effect disappears as the duration of these ads increases. We conclude with a theoretical prediction regarding the circumstances under which the display advertising industry would benefit if it moved to a partially or fully time-based standard.

Categories and Subject Descriptors: J.4 [Social and Behavioral Sciences]: Economics

General Terms: Economics, Experimentation

Additional Key Words and Phrases: display, advertising, memory, recall, recognition, exposure, time

1. INTRODUCTION

Display advertising is a \$10 billion dollar per year industry [PricewaterhouseCoopers 2011] in which advertisers pay publishers to place ads alongside content on publisher websites. Display ads are typically sold on a per impression basis, where one impression is simply one download of an ad. Furthermore, agreements between publishers and advertisers are often made through guaranteed contracts. For example, a publisher might agree to deliver 10 million impressions to men age 50–70 on finance-related pages, or 8 million impressions to people interested in sports. Display ads are also sold on exchanges like Google’s DoubleClick exchange (GDC) or Yahoo!’s Right Media Exchange (RMX). In this context, a publisher would auction the right to show a display ad targeted to a specific user in real time. Whether display ads are sold via contract or via an exchange, they are usually sold by the impression.

In addition to increasing short-term sales [Lewis and Reiley 2011; Manchanda et al. 2006], advertisers seek to increase brand recognition and brand awareness with display ads [Drèze and Hussherr 2003]. Accordingly, advertisers often measure the effectiveness of brand advertising using memory metrics [Wells 2000], which are proxies for ad effectiveness that have been in use for nearly a century [Starch 1923]. Our previous work established a causal link between the amount of time an ad is in view and the probability that a user will recall or recognize it [Goldstein et al. 2011]. In ad-

Author’s addresses: D. G. Goldstein and S. Suri, Yahoo! Research, New York, NY; R. P. McAfee, Yahoo! Research, Burbank, CA. {dan@dangoldstein.com, preston@mcafee.cc, sidsuri@ymail.com}.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

EC’12, June 4–8, 2012, Valencia, Spain.

Copyright 2012 ACM 978-1-4503-1415-2/12/06...\$10.00.

dition, this research found that there are diminishing returns to increased exposure time. That is, there was a steep increase in the probability of remembering for exposure times up to roughly 40 seconds, followed by a less steep, although still increasing, effect of time beyond that. These results suggest that time of exposure combined with the number of impressions delivered may be a better standard for pricing ads since it causally influences and more exactly captures the recognition and recall that display advertisers seek.

In the present work, we investigate how a publisher might sell time-based ads to increase the total effect on memory per unit of time, a topic of interest to both advertisers and publishers. The central question we address is how advertisers and publishers should set the duration and scheduling of display ads. Specifically, when scheduling a fixed amount of time, we ask if it is better to show multiple ads of shorter duration or fewer ads of longer duration. Longer duration ads might increase the likelihood that a user sees each ad, however, a sequence of shorter ads gives users more ads to notice. It is not *a priori* clear which would result in more overall recollection.

To answer this question we conduct an online behavioral experiment in which participants are asked to read an online news article while ads are displayed according to one of two randomly-assigned schedules. In one condition, one display ad is shown for t seconds and then replaced by another display ad for t seconds. In the other condition, a sole display ad is shown for $2t$ seconds. Two phenomena emerge. First, we find that two shorter ads result in greater recollection than one ad of twice the duration. Second, we find that this effect disappears as the durations of the ads, t , increases. We conclude by providing theoretical conditions under which the display advertising industry as a whole would benefit if time-based pricing were adopted.

1.1. Problem Definition

We want to assess the value of switching from an impression-based advertising system to a time-based advertising system. Because impressions last for a variable length of time, there is an inherent randomness in impression-based systems, in which some impressions last a few minutes while others last a few seconds; it may be useful to divide longer impressions across several ads.

To assess the value of doing so, we compare an ad of long duration to two shorter ads. Consider two impressions, shown to two users u_1 and u_2 , which last for $2t$ seconds each. Suppose that, under impression-based advertising, u_1 sees ad A and u_2 sees ad B . An alternative that becomes feasible with time-based advertising is for u_1 to see ad A for t seconds followed by ad B for t seconds, and u_2 to see ad B for t seconds followed by ad A for t seconds. The purpose of this paper is to discover which of these two schedules results in a higher probability of remembering the ad. More formally, we will compare:

$$\begin{aligned} & \Pr(u_1 \text{ remembers } A \mid A \text{ shown for } 2t \text{ seconds}) + \\ & \Pr(u_2 \text{ remembers } A \mid B \text{ shown for } 2t \text{ seconds}) \end{aligned} \quad (1)$$

with:

$$\begin{aligned} & \Pr(u_1 \text{ remembers } A \mid A \text{ shown for } t \text{ seconds followed by } B \text{ for } t \text{ seconds}) + \\ & \Pr(u_2 \text{ remembers } A \mid B \text{ shown for } t \text{ seconds followed by } A \text{ for } t \text{ seconds}). \end{aligned} \quad (2)$$

In the next section we precisely define the standard industry metrics that we use to measure memory. Observe that this comparison holds the total time the two ads are in view constant. It also holds the within-impression timing constant. That is, in both schedules each ad is shown from 0 to t seconds, and from t to $2t$ seconds exactly once. The second term in Equation 1 may be considered the false positive rate where one

claims to remember ads that were not displayed perhaps due to having seen the ad elsewhere.

The value inherent in the comparison of Equation 1 and Equation 2 is the value of showing shorter duration ads to more users, which becomes feasible under time-based scheduling. For example, consider two campaigns, each showing ten million impressions. These impressions can be approximately divided in half where one half of the impressions show ad *A* first and ad *B* second, and the other half of the impressions show the ads in the reverse order. Some of the impressions will be of too short a duration for the user to see the second ad, thus no value will be created. But the short impressions, under a time-based system, are the same as under an impression-based system, so the comparison is moot. For the longer impressions, the comparison of Equation 1 and Equation 2 accurately captures the increase, if any, from a greater number of shorter exposures. Note that it is not obvious whether the schedule underlying Equation 1 or 2 is superior. Longer durations with a single ad increase the probability that one ad is remembered, while shorter durations with two ads provides two opportunities to form memory traces.

We will also investigate, theoretically, whether a time-based system increases or decreases industry revenue. We posit two relevant effects arising from using a time-based system. First, the value to advertisers is more accurately monitored, so that high value advertisers buy more time, and low value advertisers buy less. Second, there is a change in the total awareness or recall provided to advertisers, working like a quantity increase. We provide sufficient conditions when these two effects work together to increase industry revenue.

2. RELATED WORK

Before we discuss how our work relates to the literature, we define and motivate the metrics we use to gauge the effectiveness of a display advertisement. *Unaided recall* is the proportion of site viewers who report remembering an advertiser with only a minimal prompt such as “Which advertisers, if any, did you remember being present on the website?” *Recognition* metrics use probes. *Text recognition* uses the name of the advertiser as a probe, for example “Did you see a Netflix ad on the previous page?” *Visual recognition* uses an image of an ad as the probe, asking, for example, “Did you see this ad on the previous page?” followed by an image of an ad. A vast literature in Marketing studies the effectiveness of television advertising using these and related measures [Lodish et al. 1995]. Drèze and Hussherr [2003] conducted a study on banner advertisements where they advocate the use of these memory metrics in online advertising. When referring to generally affecting the memory of an ad, whether it be recall or recognition, we will use the term “recollection”. Surprisingly few studies, however, have considered the effectiveness of *online* advertising in improving recollection. We will describe the most relevant of these next.

In our previous work we showed that exposure time has a causal effect on memory for an ad [Goldstein et al. 2011], whereas prior work had established only a correlation (see Danaher and Mullarkey [2003] and commentary in Goldstein et al. [2011]). Moreover, we showed that there are diminishing returns to this effect. The first seconds of exposure caused a steep increase in the memory for an ad, and further exposure time had a smaller, albeit still increasing impact on recollection. The implication is that, given advertisers who value memory of their ads, time of exposure is a more exact measure and thus a more efficient basis for pricing: charging based on what advertisers value allows for price discrimination and efficient allocation of advertising slots. While that work laid the groundwork for this research, it gave no guidance to advertisers and publishers as to how ads should be scheduled, which is the topic of this work.

More specifically, we seek to understand how to allocate time slots to influence overall recollection.

Sahni [2011] conducted a field experiment on a restaurant search website. The design of the experiment allowed for exogenous variation in the number and frequency of sponsored search ads users saw on the site over the course of two and a half months. The author did not study the exposure time of ads, but rather the amount of time between exposures. The key result of Sahni’s work is that increasing the time between exposures, up to two weeks, increases the probability of a purchasing event. Thus, this result may be viewed as a complement to ours. We investigate how the length and timing of exposures influences recollection, whereas Sahni showed how the amount of time *between* exposures influences purchasing. Braun and Moe [2011] used data on impressions and conversions at the individual level to construct a model of, among other phenomena, ad wearout and how different creatives should be rotated. While their research is centered on the tailoring of impressions, ours is focused on what happens second by second within an impression.

3. METHODS

The experiments reported here were conducted on Amazon Mechanical Turk¹. Mechanical Turk is an online labor market where requesters can post jobs and workers can choose which jobs to do for pay. After a worker submits a job, the requester can either accept or reject the work based on its quality. The fraction of jobs that a worker submits which are accepted is that worker’s approval rating which functions as a reputation mechanism used to help ensure work quality. Mechanical Turk was originally built to accomplish tasks that are easy for humans but hard for machines like image recognition, audio transcription and adult content classification. Hence jobs on Mechanical Turk are called “Human Intelligence Tasks” or “HITs”. There is a burgeoning literature in the academic community around using Mechanical Turk as a platform for online behavioral experiments [Mason and Watts 2009; Shaw et al. 2011; Rand 2011; Mason and Suri 2012]. In this setting, experimenters take on the role of requesters and post their experiment as a HIT and workers are the paid participants in the experiment. Recent studies show behavior observed in Mechanical Turk experiments matches behavior observed in university lab experiments extremely well [Paoletti et al. 2010; Horton et al. 2011; Suri and Watts 2011].

We used the Mechanical Turk API to restrict our participant pool to workers in the United States to help ensure that they can read and understand English. We also restricted to those workers who have an approval rating of 90% or more. The Amazon API gives each worker account a unique, anonymous identifier. By storing these WorkerIDs we were able to ensure that a worker could only do the experiment one time. In all, 1,100 participants took part in these experiments, which ran over the course of two periods of one week each. We, next describe the format of the experiment and the various treatments to which the participants were randomly assigned.

3.1. Experimental Design

Participants were paid a 50 cent flat rate for the HIT plus 10 cents for each question answered. We chose not to pay based on the correctness of the answers to alleviate incentives for sharing answers between workers. The preview page of the HIT consisted of a brief consent form along with the instructions indicating that the HIT involved reading a web page and answering questions about it.

After reading and accepting the instructions participants were then shown an image of a webpage from an actual Yahoo! website. An image was used so that if a worker

¹<http://www.mturk.com>



Fig. 1. Screenshot of the article with the Netflix ad.

clicked on what appeared to be a hyperlink, nothing would happen. The article image consisted of text with images along with a display ad. See Figure 1 for a screenshot. Since 99% of screens on the Web can show an image of 600 pixels in height² we chose this to be the height of the article image to ensure that the article and display ad were always in view in their entirety and the user never had to scroll to see any part of them. Even if a small number of users did have to scroll to see part of the page, due to random assignment, they would be evenly spread among our treatments and thus not bias our results.

The goal of this research is to compare the recollection of two short ads to one longer one. Thus, for each memory metric we compared the sum of the metric over the two short ads to the metric measured on one long ad plus the false positive rate. The treatments with two short ads necessarily involve two advertisements, and of course different advertisements may be differentially memorable. To hold all of this constant, the simplest test involves two orderings of the short ads. Denote the treatment that shows ad A followed by ad B as AB. Then the simplest test is to compare the effectiveness of AB for one user and BA for a second user, to AA for a third user and BB for a fourth. The AB/BA treatment shows each ad for the same amount of time and in the portions of an impression as the AA/BB treatment, and thus has dedicated the same amount of resources to each advertiser. If the total effectiveness of AB/BA exceeds AA/BB, then publishers should split time slots between two advertisers.

On the experimental webpage, all objects were static except for the display ads, which were changed in two different ways. In one class of treatments, an ad was displayed for t seconds, then replaced by another ad for t seconds, which was then replaced by whitespace. In a second class of treatments, an ad was displayed for $2t$ seconds and then replaced with whitespace. This design allowed us to compare the memory of two ads shown for t seconds each to the memory of one ad shown for $2t$ seconds, as described above. We used one pair of time treatments where $t = 10$ and another pair of

²http://www.w3schools.com/browsers/browsers_display.asp

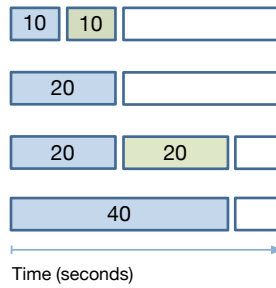


Fig. 2. The four time treatments in which participants were randomly placed. Each colored rectangle represents an ad with the number of seconds it was in view. The white rectangles on the right side of the figure indicate the absence of an ad.

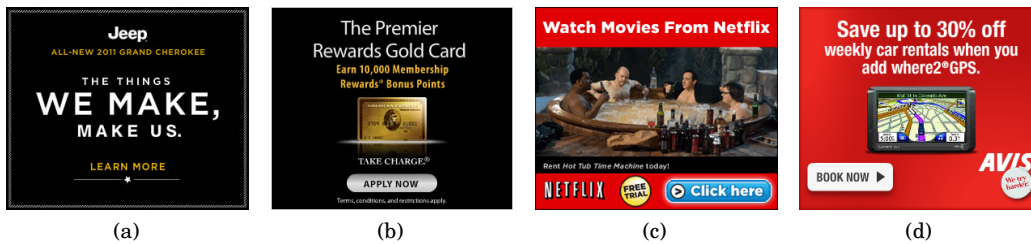


Fig. 3. The ads used as targets and lures in the experiments. When the Jeep ad (3(a)) was the target, the American Express (3(b)) ad was the lure and *vice versa*. Similarly, when the Netflix ad (3(c)) was the target the Avis ad (3(d)) was the lure and *vice versa*.

time treatments where $t = 20$ resulting in four time treatments overall. Figure 2 gives a pictorial representation of the time treatments.

We define the following notational convenience to describe the experimental treatments.

Definition 3.1. If X, Y, Z are (not necessarily different) ads and u is a user, let $\Pr(X|YZ) = \Pr(u \text{ remembers } X \mid Y \text{ shown for } t \text{ seconds followed by } Z \text{ for } t \text{ seconds})$.

If $Y = Z$ then the above definition describes a situation where the same ad is shown continuously for $2t$ seconds. Thus our experiment will measure and compare $\Pr(A|AA) + \Pr(A|BB)$ with $\Pr(A|AB) + \Pr(A|BA)$.

In the interests of external validity we used four different ads as stimuli. The first ad treatment had a Netflix ad shown first and a Jeep ad shown second; the second ad treatment had the opposite order. The third ad treatment had an Avis ad shown first and an American Express ad shown second, and the fourth ad treatment had the opposite order. In “single ad” time treatments, in which only one ad was shown, the second ads in the order described above were left out. See Figure 3 for pictures of the ads. In all, the four time treatments and the four ad treatments yielded a 4×4 , between subjects design. Subjects were randomly placed into one of these 16 treatments at the point of accepting the HIT to avoid any confound between dropping out of the experiment and the treatment assigned.

After participants finished reading the article at their own pace they clicked a link and were taken to a page where they played a game for a fixed amount of time. We chose Tetris as it is a visual game that should not create ad-specific linguistic memory interference. The game was rendered in black and white to avoid interference with the colors in the ads. The game time was chosen such that, on average, the amount of

time between the first ad disappearing and the following questionnaire was the same across conditions. This ensures that on average, each participant experienced roughly the same amount of forgetting time between the initial ad exposure and test. After playing for the designated amount of time, participants were automatically directed to a questionnaire.

Once having arrived at the questionnaire, participants were unable to press the “back” button on their browser to return to the article. Participants were asked two multiple choice reading comprehension questions about the article on the previous page, after which they were asked an unaided recall question: “Which advertisements, if any, did you see on the page during this HIT? Type the name of any advertisers here if you can remember seeing their ads on the last page, or indicate that you are unable to remember any.” The next page then consisted of four separate recognition questions with textual cues of the form, “Did you see a ___ ad?” with Netflix, Jeep, Avis, and American Express being the advertisers filling in the blank. After answering these questions participants then went to a page which consisted of four separate recognition questions with cues of the form, “Did you see the following ad?” with a picture of the Netflix, Jeep, Avis, and American Express ads following each question. The ads were chosen such that each had a strong visual resemblance to another ad, exhibited in Figure 3. The Avis lure is primarily red, much like the Netflix ad, and the American Express lure is primarily black, much like the Jeep ad. Thus when the Netflix ad was shown, the Avis ad acted as its “lure” ad and *vice versa*. Similarly when the Jeep ad was shown the American Express ad acted as its lure ad and *vice versa*. The lure ads were used to check, for example, if users were simply remembering that there was a red rectangle in the top right part of the screen or if they actually noticed the ad itself and the advertiser depicted in it.

The data from a participant were encoded as 12 binary responses. The first four responses coded mentions of the two target ads and the two lures from the unaided recall question. The next four binary responses coded the recognition questions with textual cues and the final four responses coded the recognition questions with visual cues.

4. RESULTS

From the initial sample of 1,100, we excluded observations on the basis of the following criteria, which were determined before the analysis began. We excluded 42 participants who did not complete the survey, and 22 for incorrectly answering both reading comprehension questions. Participants who took fewer than 40 seconds to read the article were not exposed to the intended time treatment and resulted in 110 exclusions. In addition, we excluded participants who took over four minutes to read the article as they may have been interrupted, yielding another 10 exclusions. The remaining 916 participants make up the set we analyze.

Table I addresses the question of whether a greater total probability of remembering an ad is achieved with two ads of length t or one ad of length $2t$. For each of the three memory metrics, the sum of the metric for the two 10 second ads ($\Pr(A|AB) + \Pr(A|BA)$, $t = 10$) is higher than the sum of the metric for the single 20 second ad and the false alarm rate ($\Pr(A|AA) + \Pr(A|BB)$, $t = 10$). For example, when $t = 10$, the probability visually recognizing the ad in the AB and BA treatments sums to .58, while that of the AA and BB treatments is only .41. Text recognition and unaided recall show a similar pattern. The p-values listed are the probabilities of these data given the hypothesis that the $\Pr(A|AB) + \Pr(A|BA) < \Pr(A|AA) + \Pr(A|BB)$. Thus, when t is 10 seconds, the total amount of recollection in AB + BA treatments exceeds the total in the AA + BB treatments. The top three panels of Figure 4 graphically represent this result. Here, for all three metrics, the dotted green line indicating the sum of the memory

measure for two 10 second ads is significantly higher than the orange line indicating the sum of the memory measure for one 20 second ad plus the false alarm rate.

However, when t is 20 seconds, a very different picture emerges. As can be seen in the bottom rows of Table I and the bottom panels of Figure 4, there is no significant difference between $\Pr(A|AB) + \Pr(A|BA)$ and $\Pr(A|AA) + \Pr(A|BB)$ and the direction of the effect reverses in two of three cases. Surprisingly, the sum of two ads ($\Pr(A|AB) + \Pr(A|BA)$) at $t = 10$ is comparable to the sum of two ads at $t = 20$, differing by at most four percentage points. Thus, not only are two short (10 second) ads better than one ad of twice the duration, they are also roughly equivalent to *two* ads of twice the duration. This result is consistent with the powerful effect of ad onset time, which we address in Section 4.3.

The main conclusion of the above analysis is that if A and B are advertisers and ad slots are short (around 10 seconds), it seems that more total impact on memory is created when splitting an impression between two advertisers than giving each advertiser its own full slot. That is, in the terminology established earlier, the memory under AB + BA is greater than memory under AA + BB. However, it is in principle possible for $\Pr(A|AB) + \Pr(A|BA) > \Pr(A|AA) + \Pr(A|BB)$ to hold averaged over many different advertisers, but not for a specific advertiser. For example, advertiser A might benefit greatly from the split impressions, while advertiser B suffers slightly. To check whether this occurs in practice, we take advantage that the experimental design uses four unique advertisers, each of which can be used as a test to see whether two short ads lead to more recall than one ad of twice the duration plus the false alarm rate. The results are shown in Table II. Again, a difference between the $t = 10$ and the $t = 20$ condition emerges. In the former case, in 11 of 12 tests, $\Pr(A|AB) + \Pr(A|BA) > \Pr(A|AA) + \Pr(A|BB)$, while in the latter case, there is no clear pattern. P-values are reported in Table II, though it should be noted that the sample sizes here are four times smaller than in Table I, decreasing statistical power; all else equal, p-values increase as sample size decreases, which explains why the aggregate results for $t = 10$ in Table I have lower p-values than do those in Table II. Nonetheless, the largely consistent sign and at times sizeable absolute magnitude of the differences in the $t = 10$ conditions lend support to the conclusion that the benefits of shorter ads hold within advertisers, especially for the visual recognition task, as shown in Figure 5.

Table I. Comparison of the sums of the memory rates for two successive ads shown t seconds each ($\Pr(A|AB) + \Pr(A|BA)$) with the sum of the memory rates for one ad shown for $2t$ seconds and the false alarm rate Standard errors are given in parentheses. The p-values, rounded to two places, are bootstrapped estimates of the probability of the data given the hypothesis that $\Pr(A|AB) + \Pr(A|BA) < \Pr(A|AA) + \Pr(A|BB)$.

Condition	Measure	$\Pr(A AB)$	$\Pr(A BA)$	$\Pr(A AA)$	$\Pr(A BB)$	$\Pr(A AB) + \Pr(A BA)$	$\Pr(A AA) + \Pr(A BB)$	p-value
10+10 vs. 20+0 ($t = 10$)	Visual	.37(.03)	.21(.03)	.35(.03)	.05(.01)	.58(.04)	.41(.03)	<.001
	Text	.22(.03)	.16(.02)	.25(.03)	.03(.01)	.38(.04)	.28(.03)	.02
	Recall	.12(.02)	.09(.02)	.15(.02)	0(0)	.21(.03)	.16(.02)	.07
20+20 vs. 40+0 ($t = 20$)	Visual	.41(.03)	.17(.02)	.52(.03)	.09(.02)	.57(.04)	.61(.04)	.74
	Text	.28(.03)	.11(.02)	.37(.03)	.07(.02)	.39(.04)	.44(.04)	.83
	Recall	.21(.03)	.04(.01)	.21(.03)	.01(.01)	.25(.03)	.22(.03)	.26

Table II. Advertiser-specific variant of Table I.

Condition	Advertiser	Measure	$\Pr(A AB)$	$\Pr(A BA)$	$\Pr(A AA)$	$\Pr(A BB)$	$\Pr(A AB) + \Pr(A BA)$	$\Pr(A AA) + \Pr(A BB)$	p-value
10+10 vs. 20+0 ($t = 10$)	Avis	Visual	.24(.06)	.21(.05)	.29(.06)	.02(.02)	.45(.08)	.31(.06)	.07
		Text	.16(.05)	.12(.04)	.17(.05)	0(0)	.28(.06)	.17(.05)	.1
	AMEX	Recall	.09(.04)	.1(.04)	.08(.03)	0(0)	.19(.05)	.08(.03)	.04
		Visual	.47(.07)	.21(.05)	.38(.06)	.06(.03)	.67(.08)	.45(.07)	.02
	Netflix	Text	.19(.05)	.19(.05)	.25(.06)	.03(.02)	.38(.07)	.28(.06)	.15
		Recall	.09(.04)	.03(.02)	.13(.04)	0(0)	.12(.04)	.13(.04)	.55
20+20 vs. 40+0 ($t = 20$)	Avis	Visual	.44(.07)	.29(.06)	.45(.07)	.06(.03)	.72(.09)	.51(.07)	.03
		Text	.22(.06)	.29(.06)	.38(.06)	.08(.03)	.51(.08)	.45(.07)	.32
	Jeep	Recall	.13(.04)	.23(.06)	.25(.06)	.02(.02)	.36(.07)	.27(.06)	.17
		Visual	.35(.07)	.13(.04)	.31(.06)	.07(.03)	.47(.08)	.38(.07)	.2
	AMEX	Text	.31(.06)	.05(.03)	.2(.05)	0(0)	.36(.07)	.2(.05)	.03
		Recall	.17(.05)	.02(.02)	.16(.05)	0(0)	.19(.05)	.16(.05)	.31
20+20 vs. 40+0 ($t = 20$)	Avis	Visual	.42(.06)	.22(.05)	.5(.07)	.11(.04)	.64(.08)	.61(.08)	.41
		Text	.28(.05)	.12(.04)	.27(.06)	.06(.03)	.4(.07)	.33(.07)	.25
	AMEX	Recall	.18(.05)	.04(.02)	.12(.04)	0(0)	.22(.05)	.12(.04)	.05
		Visual	.4(.06)	.18(.05)	.42(.06)	.06(.03)	.58(.07)	.48(.07)	.17
	Netflix	Text	.26(.05)	.14(.04)	.34(.06)	.08(.04)	.4(.07)	.42(.07)	.57
		Recall	.16(.04)	.04(.02)	.19(.05)	0(0)	.2(.05)	.19(.05)	.41
Jeep	Avis	Visual	.42(.07)	.15(.06)	.6(.07)	.13(.04)	.56(.09)	.73(.09)	.92
		Text	.35(.06)	.12(.05)	.58(.08)	.11(.04)	.47(.08)	.69(.09)	.97
	AMEX	Recall	.29(.06)	.07(.04)	.33(.07)	.04(.03)	.36(.07)	.36(.08)	.5
		Visual	.39(.08)	.09(.04)	.58(.07)	.05(.03)	.48(.08)	.63(.07)	.9
	Netflix	Text	.22(.06)	.07(.04)	.35(.06)	.02(.02)	.29(.07)	.37(.07)	.78
		Recall	.22(.06)	.02(.02)	.25(.06)	0(0)	.24(.07)	.25(.06)	.57

Table III. Comparison of the products of the complements of the recall rates of two successive ads shown t seconds each, $(1 - \Pr(A|AB)) * (1 - \Pr(A|BA))$, compared to the product of complements of the recall rates of one ad shown for $2t$ seconds and the false alarm rate, $(1 - \Pr(A|AA)) * (1 - \Pr(A|BB))$. Standard errors are given in parentheses. The p-values, rounded to two places, are bootstrapped estimates of the probability of the data given the hypothesis that $(1 - \Pr(A|AB)) * (1 - \Pr(A|BA)) > (1 - \Pr(A|AA)) * (1 - \Pr(A|BB))$.

Condition	Measure	$(1 - \Pr(A AB)) * (1 - \Pr(A BA))$	$(1 - \Pr(A AA)) * (1 - \Pr(A BB))$	p-value
10+10 vs. 20+0 ($t = 10$)	Visual	.5(.03)	.61(.03)	<.001
	Text	.66(.03)	.73(.03)	.03
	Recall	.8(.03)	.84(.02)	.10
20+20 vs. 40+0 ($t = 20$)	Visual	.49(.03)	.44(.03)	.90
	Text	.64(.03)	.58(.03)	.90
	Recall	.76(.03)	.78(.03)	.31

4.1. Remembering at Least One Ad

An alternate means for comparing the AB + BA schedule with the AA + BB schedule is to calculate, for each schedule, the probability of not remembering either of the ads. More formally, in this analysis we will compare $(1 - \Pr(A|AB))(1 - \Pr(A|BA))$ with $(1 - \Pr(A|AA))(1 - \Pr(A|BB))$. Consistent with the results thus far, Table III shows that on the measure of not remembering any ads, there is an advantage to the AB + BA schedule over the AA + BB schedule when $t = 10$. In other words one is more likely to remember at least one ad in the AB + BA schedule than in the AA + BB schedule. Also consistent with results thus far, when $t = 20$ seconds the effect disappears. This analysis supports the idea that, for relatively short time slots, there is an advantage to showing multiple ads over a single ad on the metric of remembering at least one ad.

4.2. Lure Ads

Across all conditions, the rate of incorrectly indicating memory for one of the visually similar lure ads were low and quite similar to those in [Goldstein et al. 2011]: 0% for recall, 6.6% for text recognition and 7.5% for visual recognition questions. These are similar to the false alarm rates, $\Pr(A|BB)$, shown in Table I. Thus if we asked a user if they were remember an ad that was not shown, whether or not that ad was visually similar to the ad that was shown did not have a substantial effect on the false recall rate. We can conclude that in the text and visual recognition numbers, i.e. $\Pr(A|AB)$, $\Pr(A|BA)$, and $\Pr(A|AA)$, reported in Table I, users remembered more than just the predominant color of the ad. This is most strongly supported by the unaided recall questions in which false alarms are highly improbable.

4.3. Effect of Onset Time

We define an ad's onset time to be the amount of time between the page loading and the ad appearing. If an ad appears as the page loads, it has an onset time of 0. In Table I if one compares $\Pr(A|AB)$ with $\Pr(A|BA)$ one will see $\Pr(A|AB) > \Pr(A|BA)$ for all treatments. That is, the second ad presented is at a disadvantage compared to the first. Moreover, the difference $\Pr(A|AB) - \Pr(A|BA)$ is far larger when $t = 20$ seconds than when $t = 10$ seconds. This is also visually apparent in Figures 4 and 5 if one observes that the blue lines (indicating $\Pr(A|AB)$) are always higher than the magenta lines (indicating $\Pr(A|BA)$). In addition, Figure 4 shows the difference in heights is much greater in the bottom three panels ($t = 20$) than the top three panels ($t = 10$). This shows that the longer the onset time of an ad, the less likely it is to be remembered. Accordingly, advertisers should value ads with early onset times. Onset time can also explain why two short ads are roughly as effective as two longer ads, as mentioned in Section 4. When the ads are longer, the second ad appears at a greater onset time

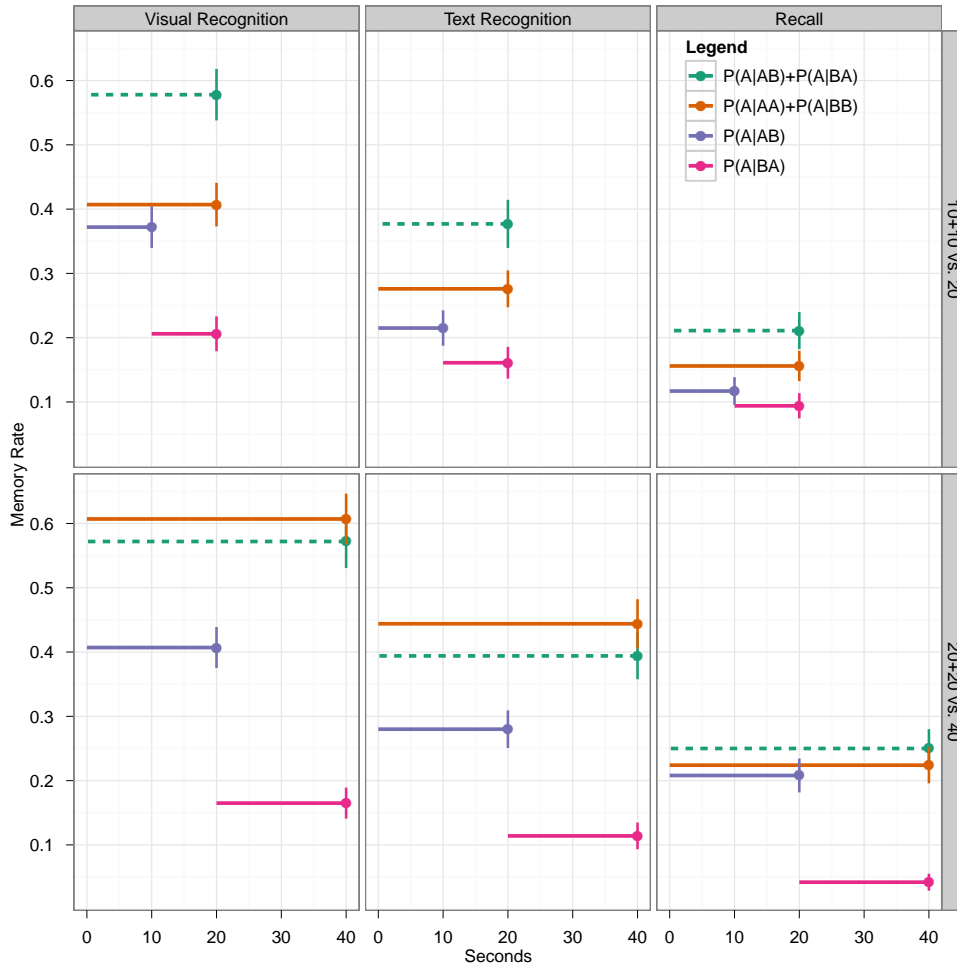


Fig. 4. The x-coordinate of the leftmost endpoint of each horizontal line indicates the time at which the ad appeared (either at 0, 10, or 20 seconds). The length of each line indicates the duration of the ad. The y-axis indicates the probability (or sum of probabilities) of remembering according to the three memory metrics. Vertical line segments are confidence intervals of one standard error. For each memory metric, the dotted green line shows the sum of the metric for the two short ads. The top three panels, the ($t = 10$) condition, compares the sum of two 10 second ads ($P(A|AB) + P(A|BA)$) to the sum of one 20 second ad and the false alarm rate ($P(A|AA) + P(A|BB)$). The bottom three panels, the ($t = 20$) condition, are analogous except that the shorter ads had a duration of 20 seconds each and the longer ads had a duration of 40 seconds each.

reducing the chance of it being remembered. In addition, this suggests that slotting more than two ads into an impression is not likely to be effective. Given ads of even short durations, such as 10 seconds, ads beyond the second would have onset times so great as to diminish their memory rates.

4.4. Do Two Ads Gain an Advantage from Apparent Motion When Loading?

As seen in Tables I & II and Figures 4 & 5, for shorter ad durations, it seems to be the case that two short ads (of 10 seconds) leads to more recollection than one longer ad of 20 seconds. A proposed explanation for this is that apparent motion of the second ad's

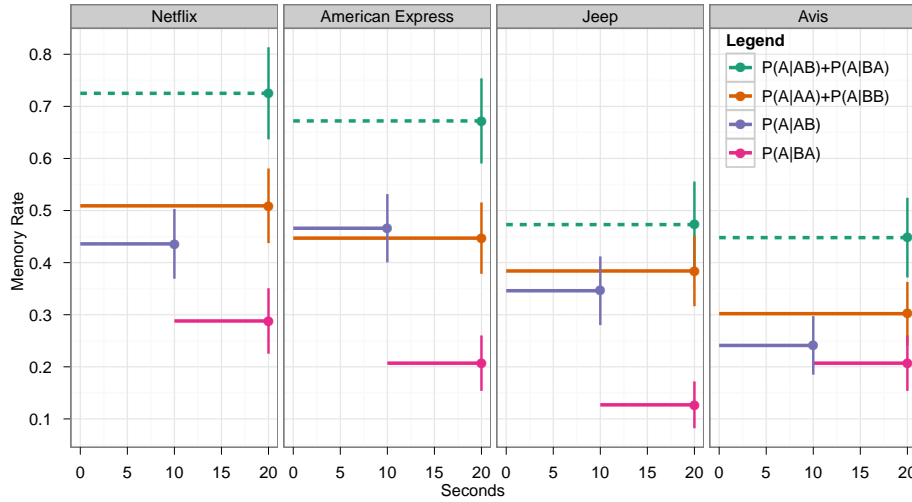


Fig. 5. The visual recognition rate only for each of the four advertisers in the ($t = 10$) condition. The x-coordinate of the leftmost point of each horizontal line indicates the time at which the ad appeared. The length of the line indicates the duration the ad remained in view (i.e. the exposure time). The y-axis indicates the probability of remembering the ad in the visual recognition test. The vertical line segments are confidence intervals of one standard error. The dotted green line shows the sum of the recognition rates of the two 10 second ads ($P(A|AB) + P(A|BA)$). The solid brown line shows the sum of one 20 second ad and the false alarm rate ($P(A|AA) + P(A|BB)$).

loading might capture user attention and provide another opportunity for the ad to be noticed. This idea has support in previous research, which found that animated ads are more likely to be noticed than static ones [Yoo and Kim 2005]. To test this explanation, we ran a follow-up study in which we replicated the condition in which the Netflix ad was shown for 20 seconds but introduced apparent motion in the ad. We did so by displaying the ad for 10 seconds, then by covering it with whitespace for 0.125 seconds, and then displaying the ad again for another 9.875 seconds. This provided the user two opportunities to detect a change, one when the whitespace appeared and one when the ad reappeared. To provide comparable statistical power to the original condition, an approximately equal number of participants was assigned to this variant as to the original. Despite this manipulation, the 20 second ad that flashed after 10 seconds was not significantly more memorable than the static version. This relationship held across visual recognition (25/56 participants recognizing in the original vs. 24/54 in the variant, p-value ≈ 1), text recognition (21/56 vs. 25/54, p-value=.46) and recall (14/56 vs. 21/54, p-value=.17). While apparent motion may have an effect at the margin, it does not seem to be the primary reason why two short ads lead to more recollection than one longer ad.

5. EQUILIBRIUM EFFECTS OF TIME-BASED PRICING

Will industry revenue be higher using impression-based pricing, or using time-based pricing? We answer this question by analyzing the market equilibrium. Selling advertisements by the page view or impression provides a noisier measure of advertiser value than selling by time, because time more closely correlates with advertiser value (recall and recognition) than impressions. In this section, we consider whether using the more accurate time-based measurement induces higher industry revenue.

Let $\nu(x)$ be the dollar value for an advertiser created from advertisement exposure x . The function ν is assumed to be increasing and concave. To illustrate ν , suppose that at each moment in time t , there is a probability r_t that an advertisement goes into a user's memory. The probability that the ad is not recalled, denoted $(1 - R)$ is just the product of the probabilities of the ad not going into memory at each moment in time, or

$$1 - R = \prod_{t=1}^T (1 - r_t) = \exp \left(\sum_{t=1}^T \log(1 - r_t) \right).$$

Taking the continuous limit, we obtain a value function which is the value of recall ρ times the probability of recall, or

$$\nu(x) = \rho R = \rho \left(1 - \int_0^x \log(1 - r(s)) ds \right).$$

This particular functional form can accommodate our major findings—that the likelihood of recall is concave in time, that starting a second ad part way through a time interval produces a greater total likelihood of recall compared to showing a single ad, and that the effect of either switching ads or of increasing duration after 20 seconds is small. The needed properties of r for these findings are that it decreases over time and ν' becomes small sometime around 20 seconds. For the analysis in this section, we will not need to assume any properties on ν beyond the assumption that ν is increasing and concave.

We model impression-based advertising as producing a random exposure length, compared to a regime of fixed exposure lengths. This is an imperfect fit to the empirical work, since neither impressions nor time is a perfect measure of advertiser value, but the thought experiment is instructive nonetheless because impression-based advertising seems to embody more randomness than time-based advertising.

To compare the random (impression-based) to the fixed (time-based) outcomes, first imagine a random allocation X ; an advertiser can buy a quantity s of X at price p . In this case, the advertiser purchasing a quantity s obtains utility $u(s) = E\nu(sX) - ps$. Thus the advertiser chooses s^* satisfying

$$0 = u'(s^*) = E\nu'(s^*X)X - p.$$

This outcome compares to the non-random allocation where a buyer facing price p per unit will maximize $\nu(x) - px$, and choose a quantity x^* satisfying $\nu'(x^*) - p = 0$. This outcome arises from our stochastic case when $X = 1$. It will be useful later to note that, in the non-random case, $\frac{dx^*}{dp} = \frac{1}{\nu''(x^*)}$.

The payment by the advertiser is $R = ps^* = E\nu'(s^*X)s^*X$. In contrast, when the allocation is not random, the agent will buy Es^*X , the total quantity available. Note, then that utility maximization implies a price of $\nu'(s^*EX)$, yielding a revenue of $\nu'(s^*EX)s^*EX$. Whether or not the quantity is random, the average quantity per buyer should be the same with both a random and non-random allocation; if in either case there were unsold units on average, the price should drop to clear the market. Thus, revenue is higher under a non-random allocation whenever $E\nu'(s^*X)s^*X \leq \nu'(s^*EX)s^*EX$.

THEOREM 5.1. *Suppose $\nu'''(x)x + 2\nu''(x) < 0$. Then payments by advertisers are higher under deterministic allocations than under random allocations.*

PROOF. Let $f(x) = \nu'(x)x$. Then $f'(x) = \nu''(x)x + \nu'(x)$ and $f''(x) = \nu'''(x)x + 2\nu''(x) < 0$ by hypothesis. Thus f is concave and thus $Ef(X) \leq f(EX)$ by Jensen's Inequality. Hence

$$R = E\nu'(s^*X)s^*X = Ef(X) < f(EX) = \nu'(s^*EX)s^*EX.$$

□

In the following corollary, Y is a *garbling* of X if $Y = X + \delta$ where X and δ are independently distributed.

COROLLARY 5.2. *If $\nu'''(x)x + 2\nu''(x) < 0$ any garbling of an allocation reduces industry revenue.*

PROOF. Immediate consequence of the concavity of $\nu'(x)x$ and Blackwell's Theorem (presented in [Marschak and Miyasawa 1968]). □

An implication of the corollary is that, if an impression-based allocation represents a more random allocation in Blackwell's sense than a time-based allocation, and the condition holds, then a time-based allocation will have a higher industry revenue.

How likely is the condition $\nu'''(x)x + 2\nu''(x) < 0$? To answer this, we consider the elasticity of demand. Recall the elasticity of demand, which measures the responsiveness of quantity to price, is given by

$$\epsilon = -\frac{p}{x^*} \frac{dx^*}{dp} = -\frac{\nu'(x^*)}{x^*\nu''(x^*)} > 0.$$

Here we consider two properties. First, demand is elastic if an increase in price reduces revenue, which occurs when $\epsilon > 1$. Second, Alfred Marshall, the inventor of the supply and demand graph as well as the concept of elasticity, famously codified properties of demand as laws. Marshall's first law is that an increase in price reduces the quantity demanded. His second, much more obscure, law is that the elasticity of demand decreases in quantity x ([Marshall 1920] at p. 102–4). Using these two properties, we can provide sufficient conditions for randomization to reduce revenues:

THEOREM 5.3. *Suppose that demand is elastic and Marshall's Second Law holds. Then $\nu'''(x)x + 2\nu''(x) < 0$.*

PROOF. By the definition of elasticity, and omitting *'s for the sake of clarity,

$$\begin{aligned} \frac{d\epsilon}{dx} &= -\left(\frac{\nu''(x)}{x\nu''(x)} - \frac{\nu'(x)}{x^2\nu''(x)} - \frac{\nu'(x)\nu'''(x)}{x\nu''(x)^2}\right) \\ &= -\frac{\nu''(x)}{x\nu''(x)} + \frac{\nu'(x)}{x^2\nu''(x)} + \frac{\nu'(x)\nu'''(x)}{x\nu''(x)^2} \\ &= -\frac{1}{x} - \frac{\epsilon}{x} - \epsilon \frac{\nu'''(x)}{\nu''(x)} \end{aligned}$$

Thus,

$$\frac{x}{\epsilon} \frac{d\epsilon}{dx} = -\frac{1}{\epsilon} - 1 - \frac{x\nu'''(x)}{\nu''(x)},$$

or,

$$\frac{x\nu'''(x)}{\nu''(x)} = -\frac{1}{\epsilon} - 1 - \frac{x}{\epsilon} \frac{d\epsilon}{dx}.$$

Because $\nu''(x) < 0$, $\nu'''(x)x + 2\nu''(x) < 0$ if and only if $\frac{x\nu'''(x)}{\nu''(x)} > -2$ if and only if $\frac{1}{\epsilon} + \frac{x}{\epsilon} \frac{d\epsilon}{dx} < 1$. Elastic demand ensures $\frac{1}{\epsilon} < 1$; Marshall's second law guarantees $\frac{d\epsilon}{dx} < 0$. □

We illustrate the concepts with four functional forms.

Example 1. $\nu(x) = x^\alpha$ for $0 < \alpha < 1$. Then $2\nu''(x) + x\nu'''(x) = 2\alpha^2(\alpha - 1)x^{\alpha-2} < 0$ and thus the addition of noise never improves industry revenue. Note, this is the constant elasticity of demand case, with demand elasticity $\frac{1}{1-\alpha} > 1$.

Example 2. $\nu(x) = 1 - (1 - x)^\alpha$ for $\alpha > 1$. This case is valid for $0 < x < 1$. Then

$$2\nu''(x) + x\nu'''(x) = \alpha(\alpha - 1)(1 - x)^{\alpha-3}(-2(1 - x) + x(\alpha - 2)) = \alpha(\alpha - 1)(1 - x)^{\alpha-3}(-2 + x\alpha),$$

So that the addition of a small amount of noise decreases revenue when $\alpha x < 2$. In particular, the condition is met with linear demand ($\alpha = 2$) for any x , but is increasingly possible as α rises.

Example 3. $\nu(x) = 1 - e^{-\alpha x}$. This case is valid when $0 < x < \infty$. In this case $2\nu''(x) + x\nu'''(x) = \alpha^2 e^{-\alpha x}(-2 + \alpha x)$, and thus the addition of noise decreases revenue when $\alpha x < 2$.

Example 4. $\nu(x) = 1 - x^{-\alpha}$ for $0 < \alpha$. Then $2\nu''(x) + x\nu'''(x) = \alpha(\alpha + 1)x^{-\alpha-2}(-2 + 2 + \alpha) > 0$ and thus the addition of noise always increases industry revenue. However, this is a case where the monopoly problem—choose x to maximize revenue—has no solution because revenue is increasing as x decreases toward zero.

In all four examples, Marshall's second law holds. Thus any failure arises because demand is inelastic—so that a destruction of quantity increases overall revenue. In these cases, randomness works like a destruction of quantity—effectively reducing output by making it random.

6. CONCLUSION

Display ads are typically sold by impression. In this pricing scheme, a two minute impression costs the same as a two second impression, despite the former having a larger effect on memory. If the total amount of time allocated to a publisher is divided into slots, we have shown that two, short ads increases memory per slot over a single, longer duration ad when slot lengths are reasonably small (around 10 seconds). In addition to increasing the total recollection, this pricing system also increases the probability of remembering at least one ad. We have shown that advertisers should be better off under this policy in the sense their ads are more likely to be remembered. Moreover, publishers may be better off since they will experience increased revenues. In this way, the scheme with two shorter ads benefits both advertisers and publishers. This result strengthens the case for moving from an impression based pricing scheme to one that is either partially or completely based on exposure time.

Increasing online ad effectiveness will likely create additional benefits for online publishers and advertisers at the expense of offline publishers such as television. In a time-based pricing scheme advertisers would be able to improve their metrics while buying the same amount of time on the internet thereby improving advertiser returns. This improvement may induce advertisers to substitute internet advertising for some other forms of advertising, bringing more money online, and increasing online publisher revenues. Advertisers must be better off as a group, because the loss of demand for offline advertising will tend to lower the prices of offline advertising. Since the performance of offline ads is the same and there is less money spent on offline advertising, the value of offline advertising should increase. Furthermore, since online advertising must be competitive with offline advertising, online advertising performance per dollar must rise. Thus some of the value from impression splitting is captured by advertisers, and some is captured by publishers.

REFERENCES

- BRAUN, M. AND MOE, W. 2011. Online advertising campaigns: Modeling the effects of multiple ad creatives. Working Paper.
- DANAHER, P. J. AND MULLARKEY, G. W. 2003. Factors affecting online advertising recall: A study of students. *Journal of Advertising Research* 43, 252–267.
- DRÈZE, X. AND HUSSHERR, F.-X. 2003. Internet advertising: Is anybody watching? *Journal of Interactive Marketing* 17, 4.
- GOLDSTEIN, D. G., MCAFEE, R. P., AND SURI, S. 2011. The effects of exposure time on memory of display advertisements. In *Proceedings of the 12th ACM conference on Electronic commerce*. EC '11. ACM, New York, NY, USA, 49–58.
- HORTON, J. J., RAND, D. G., AND ZECKHAUSER, R. J. 2011. The online laboratory. *Experimental Economics*. To appear.
- LEWIS, R. A. AND REILEY, D. H. 2011. Does retail advertising work? measuring the effects of advertising on sales via a controlled experiment on Yahoo! Working Paper.
- LODISH, L. M., ABRAHAM, M., KALMENSEN, S., LIVELBERGER, J., LUBETKIN, B., RICHARDSON, B., AND STEVENS, M. E. 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.
- MANCHANDA, P., DUBÉ, J.-P., GOH, K. Y., AND CHINTAGUNTA, P. K. 2006. The effect of banner advertising on internet purchasing. *Journal Of Marketing Research* 43, 1, 98–108.
- MARSCHAK, J. AND MIYASAWA, K. 1968. Economic comparability of information systems. *International Economic Review* 9, 2, 137–174.
- MARSHALL, A. 1920. *Principles of Economics* 8 Ed. MacMillan and Co., Ltd., London.
- MASON, W. AND SURI, S. 2012. Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, 1–23.
- MASON, W. A. AND WATTS, D. J. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, P. Bennett, R. Chandrasekar, and L. von Ahn, Eds. ACM, 77–85.
- PAOLACCI, G., CHANDLER, J., AND IPEIROTIS, P. G. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 411–419.
- PRICEWATERHOUSECOOPERS. 2011. IAB internet advertising revenue report: 2010 full year results.
- RAND, D. G. 2011. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*. To appear.
- SAHNI, N. 2011. Effect of temporal spacing between advertising exposures: Evidence from an online field experiment. Working Paper.
- SHAW, A. D., HORTON, J. J., AND CHEN, D. L. 2011. Designing incentives for inexperienced human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. CSCW '11. ACM, New York, NY, USA, 275–284.
- STARBUCK, D. 1923. *Principles of Advertising*. A. W. Shaw Company, Chicago.
- SURI, S. AND WATTS, D. J. 2011. Cooperation and contagion in web-based, networked public goods experiments. *PLoS One* 6, 3.
- WELLS, W. 2000. Recognition, recall, and rating scales. *Journal of Advertising Research* 40, 6, 14–20.
- YOO, C. Y. AND KIM, K. 2005. Processing of animation in online banner advertising: The roles of cognitive and emotional responses. *Journal of Interactive Marketing* 19, 5, 18–34.