DANIEL G. GOLDSTEIN, SIDDHARTH SURI, R. PRESTON MCAFEE, MATTHEW EKSTRAND-ABUEG, and FERNANDO DIAZ*

Some online display advertisements are annoying. Although publishers know the payment they receive to run annoying ads, little is known about the cost that such ads incur (e.g., causing website abandonment). Across three empirical studies, the authors address two primary questions: (1) What is the economic cost of annoying ads to publishers? and (2) What is the cognitive impact of annoying ads to users? First, the authors conduct a preliminary study to identify sets of more and less annoying ads. Second, in a field experiment, they calculate the compensating differential, that is, the amount of money a publisher would need to pay users to generate the same number of impressions in the presence of annoying ads as it would generate in their absence. Third, the authors conduct a mouse-tracking study to investigate how annoying ads affect reading processes. They conclude that in plausible scenarios, the practice of running annoying ads can cost more money than it earns.

Keywords: display, advertising, online, quality, compensating differential

The Economic and Cognitive Costs of Annoying Display Advertisements

In the online display advertising industry, advertisers pay publishers (websites) to run display ads that users (website visitors) see alongside other content. Online display ads are graphic images that can vary in size, shape, animation, duration, and more. Display advertising is a large industry. In 2012, display-related ads garnered revenues of more than $12 billion in the United States, or 33% of total online advertising revenue (Interactive Advertising Bureau [IAB] 2013, p. 12). Online advertising itself brings in approximately as much revenue as broadcast television and more revenue than cable television, radio, and newspaper advertising (IAB 2013, p. 18). On mobile devices, display ads are predicted to soon overtake search ads (Gartner Inc. 2013). Many of the world's most popular web destinations, such as Google, Facebook, CNN.com, and Yahoo!, are almost entirely funded by advertising, much of it display advertising.

Online display ads are often annoying. So many people want to avoid seeing online advertisements that there have been more than 200 million downloads of one ad blocker alone.[1] From an economic perspective, these annoying display ads are interesting because they can both make and cost money for publishers. They make money directly because advertisers pay publishers to run ads. They can cost money indirectly when annoyed users abandon a site, leaving the publisher with less traffic and ultimately less advertising revenue.

The costs of annoying ads extend to publishers, advertisers, and users alike. For publishers, the sale of annoying ads can be a source of tension inside the firm between parties that are concerned with maximizing short-term sales commissions and parties concerned with maximizing long-term user engagement. Anecdotally, we have heard this tension described as the "religious war" over annoying ads. The presence of annoying ads might also signal that a publisher

*Daniel G. Goldstein is Principal Researcher (e-mail: dgg@microsoft.com), Siddharth Suri is Senior Researcher (e-mail: suri@microsoft.com), and Fernando Diaz is Senior Researcher (e-mail: fdiaz@microsoft.com), Microsoft Research, New York. R. Preston McAfee is Chief Economist, Microsoft (e-mail: mcafee@microsoft.com). Matthew Ekstrand-Abueg is a doctoral student, Northeastern University (e-mail: mattea@ccs.neu.edu). This article is an adaptation and extension of the following conference proceedings article and appears here with permission of the publishers: Goldstein, Daniel G., R. Preston McAfee, and Siddharth Suri (2013), "The Cost of Annoying Ads," *Proceedings of the 22nd International World Wide Web Conference*. The authors thank Randall A. Lewis, Justin M. Rao, and David H. Reiley for helpful conversations. Bernd Schmitt served as guest associate editor for this article.

[1]See https://adblockplus.org/en/firefox/.

ISSN: 0022-2437 (print), 1547-7193 (electronic)

DOI: 10.1509/jmr.13.0439

is desperate for business. In the case of a publisher that provides vital services, such as e-mail, such apparent desperation might cause users to switch to providers that seem to be flush with resources and, therefore, more stable.

For users, the cost of annoying ads is that they interfere with the enjoyment of the very content that brought them to the site. Annoying ads may also cause users to worry about viruses, spyware, and malware infections. In addition, being annoyed is a cost in itself.

For advertisers, annoying ads gain user attention, but there may be several downsides. The use of annoying ads may cause users to distance themselves from an advertiser's brand or to question its reputability (McCoy, Everard, and Loiacono 2008). In addition, advertisers that choose the route of annoyance may, like publishers, appear desperate. If one believes the classical economic view that advertising is effective because it signals that the advertiser has plentiful resources, desperate pleas should undermine this signal (Riley 2001). Marketing research suggests that users are less likely to remember highly annoying ads (Yoo and Kim 2005) and that actively ignored stimuli such as annoying ads are evaluated less favorably (Tavassoli 2008). In addition, the widespread use of annoying ads by competing advertisers may lower ad effectiveness for all advertisers. Furthermore, annoying ads may increase the use of ad-blocking software (Edwards, Li, and Lee 2002; Li, Edwards, and Lee 2002), which could reduce the number of publishers, leaving advertisers with fewer places to advertise.

In this article, we address two main questions pertaining to annoying display ads:

1. *What is the economic cost of annoying ads?* Annoying ads presumably create a cost for publishers arising from user abandonment, but to date, this cost has not been measured experimentally. We conduct a field experiment in an online labor market, randomly varying pay rates and the presence of annoying ads, to estimate the *compensating differential*—that is, the amount of money a publisher would need to pay users to generate the same number of impressions in the presence of annoying ads as it would generate in their absence.
2. *What is the cognitive cost of annoying ads?* In two of our studies, we measure people's accuracy in classification and reading comprehension as a function of ad annoyance. In addition, we use large-scale mouse-tracking (analysis of people's mouse movements over a web page) to better understand how annoying ads affect content consumption.

We present three studies. The first is a preparatory study that asks people to rate and comment on a representative sample of ads. The goals of this study are to generate stimuli for the two subsequent experiments and to understand the ad features—animation in particular—that users find annoying. In Experiment 1, we use these stimuli to compute the compensating differential. In Experiment 2, we investigate the cognitive impact of annoying ads by measuring mouse movements, reading comprehension scores, and task completion times. The article concludes with a discussion of the managerial implications of this research.

To begin, we review the relevant literature. Several marketing investigations have explored causes of annoyance in television advertising (Aaker and Bruzzone 1985; Bellman, Schweda, and Varan 2010). Our focus is on Internet advertising, which has received less attention. Using eye-tracking technology on participants viewing web pages, Drèze and Hussherr (2003) find that users rarely focus on advertisements, a finding that has been referred to as "banner blindness" elsewhere in the literature (Benway 1998). Burke et al. (2005) varied ad types (animated, static, or absent) and measured their effects on visual search tasks. They find that the presence of ads increases the time it takes people to conduct visual searches, with no significant difference between animated and static ads. (We note that differentiating animated and static is not necessarily the same as differentiating annoying and not annoying, and we examine the relationship in our preparatory study.) They also find that animated ads are less likely to be remembered than static ads. Yoo and Kim (2005) conducted a larger experiment in which participants were randomly exposed to web pages with ads with no animation, moderate animation, or fast animation. They find that moderate animation has a positive effect on advertisement recognition rates as well as on brand attitude measures. They also find that rapidly animated (presumably annoying) banner ads can backfire, leading to lower recognition rates and more negative attitudes toward the advertiser. This finding, combined with Burke et al.'s (2005) work, lends support to the idea that annoying ads can have negative effects not only for users and publishers but also for advertisers. In a field experiment, Goldfarb and Tucker (2011) identify two types of ads that increase self-reported purchase intentions: those that are intrusive and those that match a site's content. However, they also find that ads that have both properties reduce purchase intentions. In summary, prior research suggests that a little animation or intrusiveness may increase effectiveness, but too much can backfire. Because animation is frequently cited as a cause of annoyance in this review, in our studies we make a point of experimentally varying animation.

We compute compensating differentials using Toomim et al.'s (2011) methodology, in which experimental participants are randomly assigned tasks of varying difficulty for randomly assigned rates of pay. For example, in one study, Toomim et al. paid people to transcribe images of text on web pages that were randomly assigned to be user friendly or user unfriendly. Pay rates were also randomly assigned. By analyzing the amount of work completed in each condition, they were able to compute the compensating differential—the amount of additional money a publisher would need to pay people in the user-unfriendly condition to do as much work as people did in the user-friendly condition. In this work, we use a similar method to estimate the effect of ad quality on website abandonment.

PREPARATORY STUDY: QUALITATIVE ASSESSMENT OF AD ANNOYANCE

The preparatory study has several objectives. The first is to generate sets of annoying and nonannoying (hereinafter, "good" and "bad") ads for use in the next two studies. The second is to measure the causal impact of animation on quantitative ratings of annoyance. The third is to collect and classify qualitative data on why people find ads annoying. Experimentation took place online on Amazon.com's Mechanical Turk (MTurk) online labor market (Buhrmester, Kwang, and Gosling 2011; Horton, Rand, and Zeckhauser 2011; Mason and Suri 2012; Paolacci, Chandler, and Ipeirotis 2010); participants were paid $.25 plus a bonus of $.02 per ad rated. The task was restricted to U.S.-based partici-

pants who had at least a 95% approval rating. From 163 U.S.-based participants who began the task, we analyze the 141 participants who skipped at most 1 of the 36 questions.

At the beginning of the experiment, to familiarize participants with the range of stimuli (Parducci and Perrett 1971), we showed them 36 ads (4 ads per page over 9 pages) but did not ask them to rate the ads. The 36 ads each participant saw were selected from a pool of 144 ads that were constructed in the following manner: From an online display advertising archive,[2] 72 animated display ads were selected, 36 of which were medium rectangles (300 × 250 pixels) and 36 of which were skyscrapers (120 or 160 × 600 pixels). From each of these 72 animated ads, we created a static variant by capturing the final frame of the animation sequence. This brought the total number of ads to 144 and, importantly, created a static and animated variant of each ad so that we could later measure the causal impact of animation on annoyance. Importantly, the static variants featured the same advertiser, color scheme, and overall layout as their animated counterparts.

Participants next saw 36 ads, all of a randomly chosen shape, in random order, one per page, and rated them on a five-point scale ranging from "much less annoying than the average ad in this experiment" (1) to "much more annoying than the average ad in this experiment" (5). We chose categories relative to other ads in the experiment to prevent participants from editorializing that all ads are annoying and rating them as such. For a given ad, participants were randomly assigned to see either the static or the animated variant, meaning that participants saw a mixture of animated and static ads in the 36 ads they rated but never the animated and static variants of the same original ad. After this rating task, we presented each ad that a participant rated as annoying again, along with instructions to type a few words as to why they found the ad annoying.

The mean annoyingness rating in the experiment was 2.9 on the five-point scale. Because participants were told that a rating of 3 represented "average" annoyingness for this experiment, we conclude that the 2.9 rating reflects good aggregate calibration.

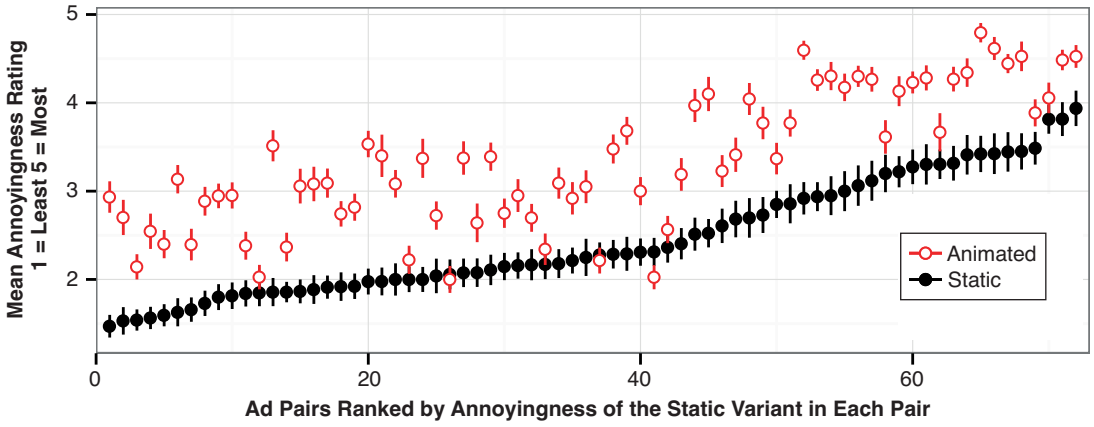
Figure 1 plots the mean annoyingness rating of the 72 ad pairs. A striking result is that animated ads were consistently rated as more annoying than static ones (mean rating 3.6 vs. 2.4; $t = 7.6$, $p < .001$), often by several standard errors. In no case was a static ad rated significantly more annoying than its animated counterpart. That is, animation seems to exert a causal effect on annoyance, holding the advertiser and ad constant. When ranking the ads, the 21 most annoying ads were all animated, and the 24 least annoying ads were all static. We designated the 10 most and least annoying ads (according to the mean ratings) as the bad and good ad sets for use in Experiments 1 and 2. Note that this implies that the bad ads are all animated and the good ads are all static.

Participants who rated an ad as annoying (a rating of 4 or 5) were asked to type reasons they found the ad annoying. They submitted 1,846 such textual responses. Taking a 5% sample of the comments, we constructed a set of high-level categories that captured the primary reasons listed for why an ad is annoying. Next, we collapsed all responses into a list of words and then counted the occurrences of each word in the list. Dropping words that occurred fewer than 10 times and "stop words" resulted in a short list of common, substantive words. We categorized each substantive word into one of the five relevant categories whenever possible. We then went back to the original long list, assigned each word of participant input to one of these five categories whenever possible, and tabulated the counts.

The most common reason given for an ad being annoying was animation. The "animation" category (typified by words such as "move," "motion," and "animate") occurred 771 times. The second category of attentional impact, which had 558 mentions, is less important for understanding what makes ads annoying because it captures the psychological impact of annoying ads (e.g., "annoying," "distracting") rather than the ad features themselves that annoy. The next most frequent category (435 mentions) was aesthetics (e.g., "ugly," "loud," "busy," "another cheap-looking ad"). A similar complaint of "poor casting or execution" was noted in Aaker

[2]See http://www.adverlicious.com.

Figure 1
MEAN ANNOYINGNESS RATINGS OF ADS



Notes: Each of 72 ads had a static and animated variant, making 144 ads. Each static ad is plotted at the same horizontal axis value as its animated counterpart, indicating that animated ads were rated as much more annoying than static ones. Error bars extend one standard error above and below the means.

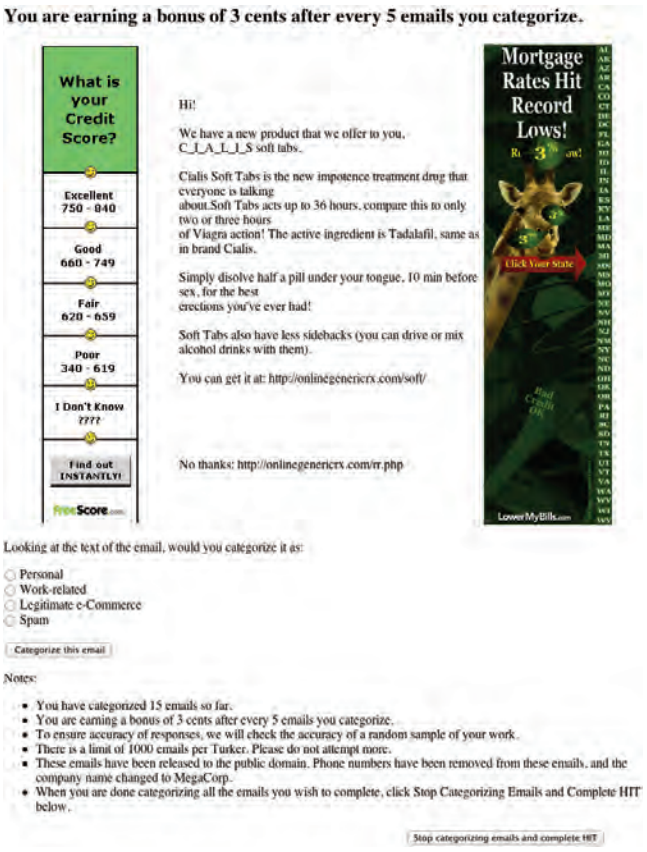
and Bruzzone's (1985) list of characteristics that make television ads annoying. For the next most often-mentioned category (122 mentions), participants used words that suggest the advertiser is disreputable (e.g., "spam," "fake," "seems like a scam"). Finally, with 107 mentions, participants expressed annoyance with the bizarre logic of the ads (e.g., "stupid," "no sense," "a dancing wizard has nothing to do with going to school"). This also corresponds to one of Aaker and Bruzzone' categories in which "the situation is contrived, phony, unbelievable, and/or overdramatized."

This preparatory study achieves two goals: (1) it provides us with sets of more and less annoying ads for Experiments 1 and 2 and (2) indicates that animation has a strong causal impact on annoyance. We do not draw causal claims about aesthetics, logic, and reputability in this experiment, because we could not vary these properties orthogonally as we did with animation. Doing so could compromise ecological validity. That is, how could one faithfully construct a Rolls Royce ad in the style of the annoying ads in Figure 2? It is not our main objective in this work to determine the drivers of annoyance, in part because this topic has been well studied in the context of television and online ads (Aaker and Bruzzone 1985; Edwards, Li, and Lee 2002; Li, Edwards, and Lee 2002). We are content to assume that where annoyance is concerned, as Supreme Court Justice Potter Stewart said of obscenity, people know it when they see it. We thus construct annoying and nonannoying ad sets on the basis of user ratings and not on ad features. Our primary focus is to understand the economic and psychological effects of annoying ads, which we turn to next.

EXPERIMENT 1: ESTIMATING THE ECONOMIC COST OF ANNOYING ADS

It has previously been shown that advertising can affect people's intent to return to a website (Li, Edwards, and Lee 2002; McCoy, Everard, and Loiacono 2008). The purpose of Experiment 1 is to measure the effect of annoying ads on website abandonment and to estimate its economic cost in the form of the compensating differential, using Toomim et al.'s (2011) method. We conducted our experiment using the MTurk labor market, with 1,223 participants with approval ratings of 90% or more. Payment was advertised as a flat rate of $.25 plus a bonus. Because the bonus was randomly assigned, it was not revealed until the task was accepted to prevent self-selection on the basis of bonus pay. The task was advertised as involving e-mail classification, which is a common task in the MTurk labor market. After accepting the task, participants were told that they would be shown e-mails and were asked to classify them as "spam," "personal," "work," or "e-commerce" related. We chose this task because of its realism and because participants could quit (i.e., stop categorizing e-mails) at any time. The number of e-mails categorized was a revealed choice and our primary dependent measure. The e-mails used in the experiment were randomly drawn from the public-domain Enron data set,[3] which provides ground-truth data regarding whether each e-mail is spam. The ground-truth data enabled us to test whether e-mail classification accuracy depended on pay rate and annoyingness of advertising.

Figure 2
E-MAIL CLASSIFICATION PAGE



Notes: The images are from the bad-ad condition with a pay rate of $.03 per five e-mails classified. Radio buttons provide available categories. Buttons allow participants to choose to quit the task or classify another e-mail. Pay rate information is displayed prominently at the top of the page. Information at the bottom of the page reiterates instructions and the number of e-mails categorized so far.

Random assignment occurred along two dimensions with three levels each, making a nine-cell experiment. One dimension was the pay rate: participants were told they would receive a bonus, per five e-mails classified, of $.01, $.02, or $.03. The other dimension determined the kind of advertising shown in the margin as people completed the task: no ads, good ads, or bad ads. The good-ad and bad-ad sets were drawn from the ten least and ten most annoying ads determined in the preparatory study. No mention was made of advertising or randomized pay conditions. The exact bonus amount was revealed to participants only after they agreed to complete the task. We ran a chi-square test to check for significant differences in the number of participants choosing to begin the task across the nine conditions and found none ($p = .25$).

In the experiment, participants were shown one e-mail per page with two good ads, two bad ads, or no ads in the margins, as in Figure 2. In the ad conditions, the two ads to the left and right of the text were randomly selected, at each page load, from the relevant set of ten good or ten bad ads

[3]See http://www.cs.cmu.edu/~enron/. We removed phone numbers, e-mail addresses, and the word "Enron" from the e-mails to safeguard privacy and to reduce distraction.

from the preparatory study. At the bottom of each page, there were radio buttons to allow the participant to classify the e-mail into the four categories and buttons either to classify another e-mail or to quit the experiment. The text width and page width were such that the page would be visible without scrolling for the majority of screen resolutions and was held constant across conditions. Ad images were named in such a way that they would not be suppressed by ad-blocking software. In the no-ad condition, white rectangles (which matched the page background) were displayed in place of the ads.

Each e-mail to be classified was shown on a new page, so viewing one e-mail constituted one impression. As soon as the task was accepted, one e-mail was presented, meaning that each participant generated at least one impression. The primary dependent measure is the number of impressions (i.e., e-mails classified) per person per condition. The mean number of e-mails classified was 61. The median was 25, and the first and third quartiles were 6 and 57, reflecting strong skewness. Only 2 of the 1,223 participants reached the upper limit of classifying 1,000 e-mails. We present means and standard errors per condition in Table 1. The randomly assigned ad quality did not affect the likelihood of classifying an e-mail as spam, which was similar across conditions (47.5%, 50%, and 48.5% in the bad-, good-, and no-ad conditions, respectively; $p = .975$ by a chi-square test).

Figure 3 shows that the difference between bad, good, and no ads stays relatively stable as outliers are removed from the distribution. In general, the data suggest that higher pay causes more impressions and bad ads cause fewer impressions. One apparent anomaly in Table 1 is that at the $.01 pay rate, the no-ad condition is lower than the ad conditions. However, at the $.01 pay rate, there is no significant difference in the number of e-mails classified according to ad condition, according to an analysis of variance (ANOVA) ($p = .42$) and inspection of the coefficients of a generalized linear model (GLM). Caution should be taken inspecting means when data are so highly skewed. To properly analyze overdispersed[4] count data such as these, a negative binomial GLM is suitable (Venables and Ripley 2002). Table 2 shows parameter estimates of two negative binomial GLMs. Because Model 1 can be difficult to interpret in log terms, Figure 4 shows Model 1's predictions on the original scale: that good ads and no ads are predicted to have a similar effect, with bad ads causing substantially fewer impressions.

[4]The ratio of the observed variance to the theoretical Poisson variance is 228.7, which suggests overdispersion ($p < .001$).

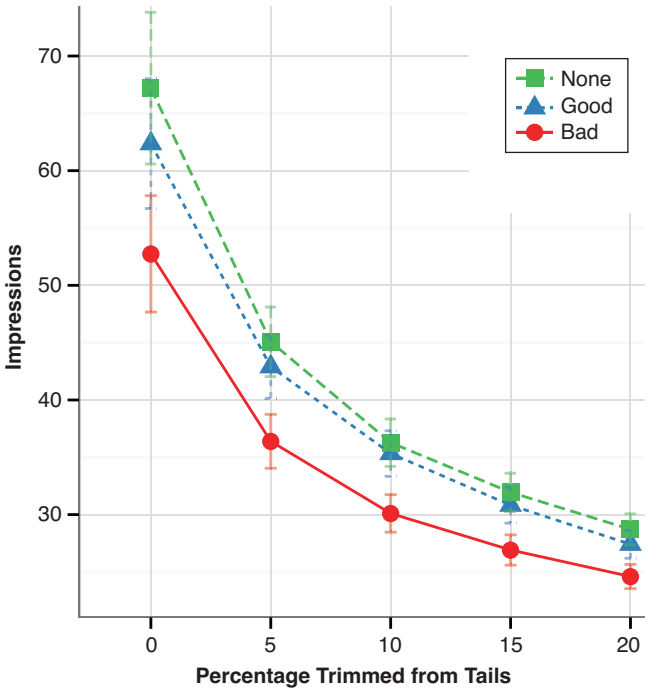
Table 1
AVERAGE NUMBER OF IMPRESSIONS BY CONDITION

Pay Rate	*Bad*	*Good*	*None*
.01	42.3 (7.2)	50.2 (9.6)	35.6 (6.8)
.02	55.9 (9.8)	55.6 (7.0)	83.2 (15.2)
.03	57.9 (8.7)	81.8 (12.6)	82.9 (10.9)

Notes: One impression is one e-mail classified. Standard errors are in parentheses.

Given the nonlinear curves in Figure 4, there are many points at which a compensating differential could be calculated. For a simple approximation, we estimate the effect of pay rates by averaging the increase in impressions related to the .2- to .4- and .4- to .6-cent pay raises. Similarly, we estimate the effect of moving from bad ads to no ads at the .4 pay rate. Doing so suggests that a .2-cent pay raise leads to an increase of 16.58 impressions and that moving from bad ads to no ads leads to an increase of 12.68 impressions.

Figure 3
ROBUSTNESS CHECK



Notes: Each data point is collapsed across pay conditions. The drop in means reflects the skewness in the data. Error bars are one standard error above and below the mean.

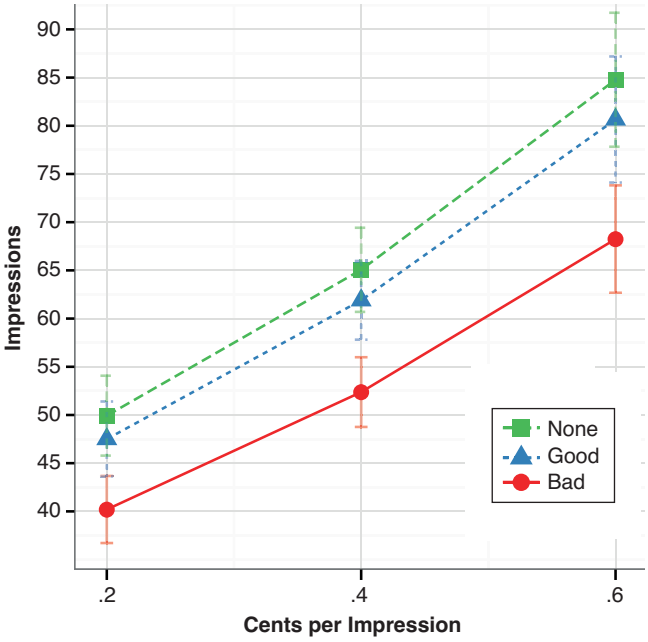
Table 2
MODELS PREDICTING NUMBER OF IMPRESSIONS

	Model 1	*Model 2*
Intercept	3.43 (.12)***	3.43 (.12)***
Good ads	.17 (.10)*	
No ads	.22 (.10)**	
Good or no ads		.19 (.08)**
Pay rate	26.47 (4.8)***	26.61 (4.8)***
Akaike information criterion	12,158.57	12,156.85
Nagelkerke pseudo R^2	.04	.04
Log-likelihood	–6,074.29	–6,074.43
Deviance	1,481.00	1,481.04
Number of observations	1,223	1,223

*$p < .1$.
**$p < .05$.
***$p < .001$.

Notes: Models are negative binomial generalized linear models. In Model 1, bad ads led to significantly fewer impressions than no ads and marginally fewer impressions than good ads. The pay rate is in dollars per five impressions, and standard errors are in parentheses. In Model 2, good ads and no ads are treated as one category.

Figure 4
IMPRESSIONS BY PAY RATE AND AD CONDITION



Notes: Impressions refers to e-mails categorized. Error bars extend one standard error above and below the predicted values.

Therefore, the pay raise required to match the effect of moving from bad ads to no ads is .153 cents per impression (.2 × 12.68/16.58). In other words, a participant in the bad-ad condition would need to be paid an additional .153 cents per impression to do as much work (i.e., generate as many page views) as a participant in the no-ad condition. In CPM (cost per thousand impressions) terms, the cost of bad ads in this experiment was $1.53 per thousand impressions. Notably, many bad ads pay less than $1.53 CPM. Indeed, recently, 53% of all display ad impressions were estimated to pay between $.10 and $.80 CPM.[5] This suggests, if the results of this experiment generalize, that bad ads actually cost publishers more money than they bring in. It also means that had we received a $.50 CPM to run annoying ads in this experiment as our e-mails were categorized, holding all else constant, it would have been better not to have run them at all.

For many major web portals, giving up advertising is not a likely option, so here we calculate the cost of bad ads relative to good ads. Moving from bad ads to good ads at the .4-cent pay rate leads to an estimated additional 9.52 impressions. Therefore, a .115-cent per impression pay raise would be required to compensate for the cost of bad ads relative to good ads (.2 × 9.52/16.58). A participant in the bad-ad condition would need to be paid $1.15 per thousand impressions to generate as many impressions as a participant in the good-ad condition. Again, if these estimates generalize, this suggests that bad ads could lose money because they typically pay publishers $.50 CPM or less. By a similar calculation, we observe that the cost of good ads relative to no ads is $.38 CPM, noting that this estimate is based on a statistically insignificant difference.

To this point, we have examined the effect of annoying ads on dropout, which is a primary concern of publishers. We turn next to the users' perspective and consider the effects of annoying ads on a cognitive task. Recall that for each e-mail classified, the Enron e-mail corpus contained ground-truth information regarding whether it was spam, which enables us to test the effect of ad types on e-mail classification accuracy. In general, classification accuracy was high at 91%. Table 3 shows the results of two regressions predicting individual accuracy rates, controlling for the number of e-mails categorized. Against a baseline of annoying ads, people classified e-mails more accurately in the presence of good ads or no ads. Because ad conditions were randomly assigned, we conclude that annoying ads have a causal impact on accuracy. The regressions imply that accuracy drops approximately 1 percentage point per 128 e-mails classified, which could reflect fatigue or a selection effect by which the set of people who decide to persist at the task are those who care less about accuracy. One concern with this regression is that there could, in principle, be another, more specific selection effect: there could be a respondent group that both persists in the presence of bad ads and does not care about accuracy. To test for this possibility, we examined accuracy binned by the number of e-mails categorized. Across all four quartiles of the distribution of e-mails categorized, the accuracy of the people assigned to the bad-ad condition was approximately 2 to 3 percentage points lower (3.1%, 2.9%, 1.9%, and 2.0% from lowest to highest quartile of e-mails categorized) than those in the no-ad condition. Similarly, a regression does not indicate a significant interaction between the bad-ad condition and the number of impressions, suggesting that the negative impact of bad ads on accuracy is stable relative to the number of e-mails classified and not due to self-selected dropout.

Thus far, we have observed that annoying ads cause people to abandon paying tasks and that these ads seem to have a negative effect on a cognitive task, namely, classifying e-mails. What we do not yet know is why. For example, annoying ads might have affected accuracy because they distracted people and impaired the reading process or

Table 3
MODELS PREDICTING CLASSIFICATION ACCURACY RATE

	Model 3	*Model 4*
Intercept	.90 (.01)***	.90 (.01)***
Impressions	–7.8e-5 (.00)**	–7.8e-5 (.00)**
Good ads	.02 (.01)*	
No ads	.03 (.01)**	
Pay rate	.14 (.43)	.14 (.43)
Good ads or no ads		.02 (.01)**
R^2	.02	.01
Number of observations	1,057	1,057

*$p < .05$.
**$p < .01$.
***$p < .001$.
Notes: Impressions refers to the number of e-mails categorized. Models exclude people who did not classify any e-mails because their accuracy rate would not be defined. For this reason, there are somewhat fewer observations than participants. The pay rate is in dollars per five impressions, and standard errors are in parentheses.

[5]See http://www.turn.com/sites/default/files/Global_Digital_Audience_Report_October_2013.pdf.

because they signaled to people that the site creators do not care about them, causing participants to take revenge by not doing careful work. In the next experiment, we introduce a task designed to gain more psychological insight into what annoying ads do to website users.

EXPERIMENT 2: THE COGNITIVE COST OF ANNOYING ADS

In the previous experiment, we find that bad ads cause participants to drop out earlier and to exhibit lower classification accuracy. The motivation behind Experiment 2 is to gain some insight into why these effects arise.

One way to understand psychological processes, especially those involving tasks done at a computer, is with eye tracking (Buscher, Cutrell, and Morris 2009). Eye-tracking studies provide fine-grained data on where the gaze of a participant is focused. The drawback of this technique is that it can be difficult to recruit participants to physically appear in a lab for such a study, often resulting in small sample sizes. Furthermore, eye-tracking equipment can be expensive. One way around these drawbacks is to use mouse tracking. Mouse tracking is known to be correlated with eye tracking, especially when it comes to measuring the number of times the eye or the mouse enters an area of interest on a computer screen and the amount of time spent in these areas (Chen, Anderson, and Sohn 2001; Guo and Agichtein 2010; Huang, White, and Buscher 2012). It is also believed to serve as a proxy for user interest (Navalpakkam and Churchill 2012; Willemsen and Johnson 2010). Because the necessary JavaScript code can be embedded into a web page, mouse-tracking studies can be conducted cheaply and at scale online (Mueller and Lockerd 2001).

From Experiment 1, the phenomena to explain are higher dropout and lower accuracy in the presence of bad ads. Because many people complained of distraction or demands on their attention in the preparatory study, we propose a "distraction hypothesis": the dropout and lower accuracy are due to annoying ads disrupting the reading process. If distraction is at play, we would expect participants to take a longer time to read a given text and to read more deliberately to compensate for the distraction. This hypothesis is based on the self-reports of distraction from the preparatory study, as well as empirical observations that distraction increases reading time (e.g., Carlson et al. 1995; Connelly, Hasher, and Zacks 1991). A few reviewers of this article believed that distraction might increase reading speed or cause less deliberate reading. To accommodate this alternative explanation, we test a more general version of the distraction hypothesis, which predicts that people simply change their reading behavior (e.g., either faster or slower, either more or less deliberately) when distracted.

To preview the results, we do not find support for the distraction hypothesis. People do attend more to bad ads than to good ads in the experiment, but reading behavior with regard to the text itself seems unchanged as a function of ad condition. Other processes may explain dropout and lower accuracy in the presence of bad ads, and we subsequently speculate about what those might be. Although this experiment does not pin down an exact mechanism behind dropout and lower accuracy, it can also be viewed as a qualitative investigation of what annoying ads do to the experience of consuming web content.

We conducted our experiments using MTurk. We restricted our participant pool to those living in the United States who had an approval rating over 97%. The first page of the experiment consisted of a consent form, a simple set of instructions, and the payment scheme. Participants were told they would read a web page and then answer a few questions. The participants were paid a $.50 (U.S.) flat rate plus $.10 per question answered. After participants read a text passage, they answered four questions. We paid per question answered as opposed to paying for correct answers to remove an incentive for participants to share answers outside the study.

After participants accepted the instructions, they were shown an image of an actual web page taken from a popular news site. We used an image of the web page (as opposed to rendered HTML) so we could ensure that the layout of the page would be uniform across all browsers and screen sizes. Rendering the article as an image also ensured that the URLs in the web page could not be followed; however, attempted clicks were recorded. The text of the article consisted of a story involving school teachers and had an accompanying graphic (see Figure 5, Panel A). Participants were randomly placed into one of three treatments: bad ads, good ads, or no ads. In the bad- and good-ad conditions, a randomly chosen bad or good ad appeared to the right of the article. The result was similar in layout to many modern web pages. In the no-ad condition, nothing was placed to the right of the article, giving it a different white-space geometry and making it suitable only for measuring page-viewing time and comprehension, but not for mouse tracking (due to "parking" effects, as we explain subsequently).

The sets of good and bad ads used in this study were five of the good and five of the bad ads used in the prior study. We exclusively used skyscraper-dimension ads to ensure that the page layout would be identical in the good- and bad-ad conditions so they could be compared directly; when page content and page layout both change in a mouse-tracking study, it is not possible to identify which of the two changes is responsible for associated changes in mousing behavior.

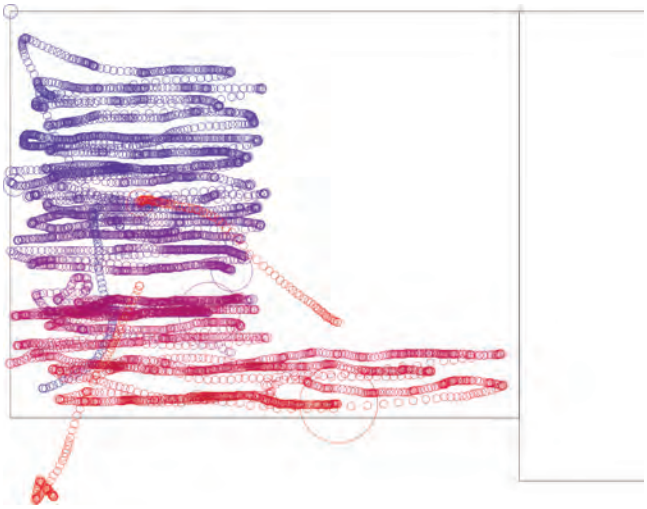
After the participants finished reading the article, they proceeded to a page on which they were presented with three multiple choice questions about what they read. Each question had five or six possible answers. We designed the study so that the answer to the last of the three questions could be found in the second-to-last sentence of the article. This arrangement tested whether a participant read to the end of the article. The fourth and final question served as a manipulation check and asked: "Was there anything on the page that make it difficult to read and understand the article? If there was nothing, then indicate that."

A total of 2,840 people completed our study, with 962, 959, and 919 participants assigned to the bad-, good-, and no-ad conditions, respectively. Participants were randomly assigned to conditions, and there were no significant difference in cell counts (chi-square, $p = .54$). As a manipulation check to ensure that the annoying ad condition was indeed annoying, we coded responses to the fourth question as to whether it referred to being annoyed by an ad. In the bad-ad condition, 41.5% of participants complained; in the good-ad condition, 4.5% complained; and in the no-ad condition, 0% complained. We thus conclude that the bad ads annoyed participants as expected.

Figure 5
HEAT MAP AND MOUSE MAP

A: Heat Map of Fixations

B: Mouse Map



Notes: Panel A shows a heat map of fixations for the bad ad condition in which red colors reflect more fixations and blue colors reflect fewer ones. Panel B presents a mouse map showing position of a participant's mouse as a web page is read. The rectangles indicate the text area and ad area. A circle is drawn each time the browser generates a mouse movement event. The size of each circle is proportional to the amount of time the mouse was left at a fixed position. The maximum circle size (straddling the bottom of the text area rectangle) indicates the mouse was held at a position for five seconds or more. The color of the circles changes from pure blue to pure red as a function of the time at which the position was recorded, relative to the total amount of time spent on the page. In both panels, the activity on the lower left-hand side corresponds to the position of the "next" button.

We begin by first determining whether the ad treatment had an effect on the amount of time participants spent on the page. Participants in the bad-ad condition spent an average of 73.1 seconds on the page, with a standard error of 1.2 seconds, while participants in the good-ad condition spent an average of 69.2 seconds, with a standard error of 1.1 seconds. An ANOVA on the log-transformed (for skewness) time data shows that this difference is statistically significant ($p = .02$). The no-ad treatment had a mean of 71.9 seconds and a standard error of 1.4 seconds. Furthermore, participants in the bad-ad condition took longer to view the page than those in the combined innocuous (good-ad and no-ad) conditions ($p = .02$, ANOVA of log time data). The difference from the good-ad condition alone was significant ($p = .02$), while the difference from the no-ad condition alone was not ($p = .11$). Note that time spent on the page is different than time spent reading the article (a key metric for the distraction hypothesis) because it may involve looking at the ad as well as at the text. We use the mouse-tracking data as a proxy for time spent attending to the ad and text separately.

A common dependent variable in eye-tracking studies is the fixation (Duchowski 2007). In this mouse-tracking study, we consider a fixation to occur when the mouse stays within a radius of 20 pixels for 300 milliseconds. Figure 5, Panel A, is a heat map of participants' aggregated fixations.

Visual inspection of the heat maps based on fixations gives the impression that there are more fixations on the ad areas when there are bad ads (relative to good ads) and that fixations on text areas seem not to be affected by ad conditions. Because heat maps are more exploratory, in what follows, we quantitatively test fixations and other measures of reading behavior.

To understand how reading behavior may change according to condition, we measure and report on several dependent variables with the mouse-tracking data. In addition to the number of fixations, we also measure the amount of time the mouse spends over the ad, the distance the mouse travels over the ad, and the number of entrances the mouse makes over the ad area. Table 4 shows all these dependent variables for both ad treatments. A consistent story emerges. Compared with good ads, bad ads cause more fixations on the ad, greater distance traveled over the ad, more entrances into the ad area, and more time spent over the ad. We tested the significance of these effects by log-transforming them and comparing the means between these two treatments. We ran simple regressions to test significance and found all the effects to be significant (see Table 4). Notably, people made 183% more ad fixations and spent 70% more time on the bad ads than the good ads. When modeling mouse movements on the basis of treatment, we find that the 41.5% who complained about the ads in the bad-ad condition showed even stronger effects compared with the good-ad condition, but we withhold these results for brevity.

Using the mouse as a proxy for attention, we find that annoying ads are noticed more than benign ads. This alone might explain why participants spent more time on the page in the presence of bad ads: it takes time to look. However, the distraction hypothesis makes a different prediction, namely, that people will read more slowly or deliberately when annoying ads are present to compensate for the distraction. Or, more generally, if people are distracted, one would expect them to somehow read differently. We address this issue next.

Focusing our analysis on the article text (as opposed to the ad), we measure the same dependent variables. As we show in the bottom rows of Table 4, and perhaps surprisingly given the ad results, there is no appreciable difference in mousing over the text according to ad condition. There is a slight difference on fixations (5% more fixations with bad ads), but this could be a false alarm given the modest p-value, the multiple comparisons, and the finding that none of the other measures differ. As a first robustness check, we collected the same measures of mouse activity over the text or ad for just the first 30 seconds after the page loads (i.e., when more than 90% of participants have yet to proceed to

Table 4
MOUSE-TRACKING METRICS

Area/Measure	*Bad Ads*		*Good Ads*		p-*Value*
Ad					
Fixations	4.45	(.67)	1.57	(.26)	<.001
Distance	182.6	(7.90)	157.9	(8.00)	.003
Entrances	1.31	(.05)	1.13	(.05)	.004
Time (milliseconds)	1,873	(321.00)	1,101	(186.00)	.001
Text					
Fixations	135.7	(9.07)	128.0	(9.05)	.047
Distance	1,492	(55.20)	1,570	(66.60)	.677
Entrances	1.51	(.06)	1.50	(.07)	.322
Time (milliseconds)	38,268	(1,207.00)	36,637	(1,085.00)	.596

Notes: Top rows: Mouse-tracking metrics from the advertisement area (rectangle to the right of the text). All these proxies for attention were significantly greater in the bad ad condition. Bottom rows: Mouse-tracking metrics from the text area. In contrast to the ad area, mousing behavior did not change substantially in the text area as a function of ad condition. The p-values are from the ordinary least squares regression coefficients. The log of each measure was regressed on the ad condition (good or bad).

the next page) and obtain the same basic results. As a second robustness check, we compared the 41.5% who were annoyed by the ad in the bad-ad condition (according to self-reports) with the good-ad condition. A similar pattern of nonsignificant differences appeared, and even the fixations measure became nonsignificant ($p = .20$), despite the large number of observations (959 in the good-ad condition, 399 in the bad-ad condition) in the regression.

As an additional check, we wanted to ensure that the apparent attention paid to the bad ad was not due to an artifact, such as people choosing to "park" (i.e., leave for five seconds or more) the mouse differently depending on whether a good or bad ad is present. Parking is a relatively common mouse behavior, but it does not represent active interest in an area. On the contrary, people tend to park the mouse on relatively *uninteresting* areas of the screen, such as the white space to the immediate right of any text or graphics. For this reason, we cannot directly compare the no-ad condition with the good- and bad-ad conditions for mouse tracking. Accordingly, we wrote a program to analyze the mouse movements and to detect incidences of parking on the text, on the ad, and to the right of the ad. There are no significant differences in the proportions of participants parking the mouse one or more times in these key areas. In the bad-ad condition, 54.2% of participants parked the mouse in the text area, compared with 56.2% in the good-ad condition ($p = .40$, chi-square test). A similar pattern held for parking on the ad (3.4% vs 3.0%, $p = .71$) or in the right margin (33.3% vs. 30.6%, $p = .23$). This result is robust to other tests, such as parking two or more times or regressing on the number of parking incidents observed. We therefore conclude that the results in Table 4 are not due to parking.

In contrast to the prediction of a distraction hypothesis, our findings indicate that the bad ads did not affect how the text was read. To probe more deeply into this question, we created "mouse maps" for each participant, as well as "mouse movies" that allow one to watch how a participant moved the mouse while reading.[6] Figure 5, Panel B, shows data from one of the experimental participants exemplifying "mouse reading"—that is, moving the mouse along various lines of text as they are read. In this experiment, as we show, approximately 5% of participants exhibited this type of mouse-reading behavior. Under the distraction hypothesis, we would expect to observe changes in the incidence of mouse reading in the presence of bad ads. Alternatively, if the bad ads were merely annoying but did not affect the reading process, which is generally consistent with the results in Table 4, the likelihood of mouse reading would be unaffected. To test this possibility, we asked three judges, blind to the conditions, to rate all 1,977 mouse maps from the bad- and good-ad conditions according to whether they represented instances of mouse reading. Pairwise agreement between the three raters was 94.3%, 95.4%, and 96.4%. Using a majority criterion to classify maps, the judges observed mouse reading in 67 of 989 (6.8%) participants in the bad-ad condition and 71 of 988 participants in the good-ad condition (7.2%), an insignificant difference ($p = .786$ according to a chi-square test). Using a unanimity criterion produced a similar result (4.8% vs 5.9%; $p = .313$ according to a chi-square test). Thus far, the mouse-tracking analysis suggests that the bad ads receive more attention than the good ads but that reading behavior is not affected by the annoyingness of ads.

Recall that we asked three reading comprehension questions. We defined an overall accuracy index, ranging from 0 to 3, which is simply the number of questions a participant answered correctly. With this measure, we observe that the bad-ad condition is significantly less accurate than the no-ad condition using the Tukey honestly significant difference test for multiple comparisons (mean difference of .076, $p = .012$) and no significant differences between the other two pairs of conditions. Similarly, the combined ad conditions were significantly less accurate than the no-ad condition (mean difference of .06, $p = .008$ according to an ANOVA). Two of the three questions were apparently easy (more than 95% correct in all treatments) and had a ceiling effect in their results. The third question, which occurred at the end of the passage, was more discriminating and showed a 6.2–percentage point difference between conditions (64.9%, 67.8%, and 71.1% in the bad-, good-, and no-ad conditions, respectively). An alternate explanation is that mouse tracking reveals whether people pay less attention to the last paragraph when bad ads are present. To test this notion, we defined a subset of the text rectangle that included only the last four lines—lines that held the answer to the third question—and measured the same metrics as in Table 4. The result is mixed. As might be expected, people in the bad-ad condition (who were more likely to get the question wrong) had fewer entrances into the relevant area than people in the good-ad condition (M = 1.95, SD = .06 vs. M = 2.18, SD = .08; $p = .004$ according to a regression on log values). They moved the mouse a shorter distance there as well (M = 451.5, SD = 17.6 vs. M = 527.8, SD = 21.1; $p = .014$). However, there was no significant difference in the number of fixations (M = 32.38, SD = 2.52 vs. M = 30.15, SD = 2.57; $p = .738$) or the mouse time in milliseconds spent in that region (M = 8,049, SD = 588 vs. M = 7,748, SD = 526; $p = .804$). Perhaps because only 5% of people read line by line with the mouse (as noted previously), mouse-tracking may be too blunt of an instrument to detect attention to very

[6]See https://archive.org/details/mousemap_1424_201407.

small areas of a page, and eye tracking would be better suited to perform such a test.

Taking these results as a whole, if mouse movements are good proxies for attention on large sections of a page, users' attention does indeed seem to be captured by annoying ads. In the ad space, time, distance, fixations, and entrances were all significantly greater for bad ads than for good ads. Furthermore, these results do not seem to be due to artifacts, such as users trying to click the ads or deciding to park the mouse differently in the presence of bad ads; there was no significant difference in click rates or parking behaviors between conditions. Users presented with bad ads took a few seconds longer to complete the task. Notably, annoying ads did not seem to affect mouse behavior, reading behavior, or time spent on the text area. To gain additional insight into the reading process, we examined the average x-coordinate of the mouse as a function of time after the page loads and noticed that users in the bad-ad condition tended to move the mouse toward the bad ad for the first 10 seconds after the page loaded and then moved it back toward the text. This tendency may suggest the bad ad is noticed just after the page loads and is less likely to be attended to as time goes on. Both findings are consistent with the overall mouse-tracking results we report herein and with recent investigations of online ad exposure (Goldstein, McAfee, and Suri 2011, 2012), which suggest that viewers typically scan the whole page just after it loads, followed by a period in which they focus mostly on the text. Such early inspection of the ad would be consistent with the idea that people in the bad-ad condition took more time to look at the annoying ad but read the text in a way that was relatively unaffected.

If annoying ads do not affect how a text is read, why was accuracy affected in Experiments 1 and 2? It might be that working to ignore ads over an extended time is somehow cognitively depleting (Gilbert, Krull, and Pelham 1988; Smit, Eling, and Coenen 2004) and leaves users with fewer resources to be accurate. To investigate this notion, we examined the data from Experiment 1 to observe whether the deleterious effect of bad ads on accuracy increased over time. However, we instead found the difference with the no-ads condition to be constant. An alternative account as to why bad ads harmed accuracy is that users expressed their dissatisfaction with the annoying ads by exerting less effort on the e-mail classification and reading comprehension questions. To satisfactorily address why annoying ads cause dropout and decreased accuracy, further research is needed.

CONCLUSIONS

Summing up the empirical work in this article, in a preparatory study we found that some ads are perceived as much more annoying than others. Among the complaints, animation was preeminent and exerted a causal effect on annoyingness ratings. Poor aesthetics and questionable advertiser reputability were also frequently mentioned. Experiment 1 shows that annoying ads can exert a causal effect on website abandonment relative to good ads or no ads. In addition, annoying ads decreased accuracy in an e-mail classification task. This experiment enabled us to estimate the compensating differential. In our study, to motivate a person to generate as many impressions in the presence of bad ads as they would in the presence of no ads or good ads, we would need to pay them roughly an additional $1 to $1.50 per thousand impressions. Experiment 2 collected process data to gain insight into how annoying ads affect content consumption. Here, we find that annoying ads garnered significantly more attention than controls, as proxied by the mouse-tracking measures of time, duration, entrances, and distance. In addition, annoying ads increased task completion time and led to slightly lower accuracy on reading comprehension questions, especially for a question referring to the end of the passage.

Publishers are often paid less than 50 cents per thousand impressions to run annoying ads, half as much as the estimated economic damage they incurred in our experiments. If the results of this experiment generalize, accepting such a low price may be a losing proposition. Whether running annoying ads indeed loses money depends on many factors, including the alternatives in the market. In our studies, participants' alternatives were finding another task on MTurk or finding something else to do on the Internet altogether. If annoying ads are running on a unique and valuable site (e.g., imagine if there were only one free e-mail provider in the world), we would expect users' tolerance for annoying ads to be high. Conversely, if annoying ads are running on a site that offers what many other sites do (e.g., news stories from mass-market newswires), switching costs are low and people's tolerance for annoying ads should be much lower. Nonetheless, managers should be able to adapt our methodology to learn about sensitivity to annoying ads on their own sites. For example, by random assignment to ad conditions, site owners could detect whether certain ads are causing abandonment and could then react accordingly, perhaps by charging advertisers for the externalities they impose.

We conclude by returning to the two main questions that motivated this work:

1. *What is the economic cost of annoying ads to publishers*? Our field study indicates that annoying ads do cause dropout and that we needed to pay people more than $1 CPM to compensate for it. In realistic settings, the practice of running annoying ads can cost more money than it earns. While web publishers do not pay users directly, the lesson should be that annoying ads will have to be compensated for somehow, such as through higher value content, to retain users. This short-term cost estimate can be understood as a lower bound on the total cost of annoying ads. There may be longer-term costs. For example, upon dropping out, users may decide never to return to a site that annoyed them.
2. *What is the cognitive impact of annoying ads*? In our studies, people seem to notice annoying ads and complain about them and were more likely to abandon sites on which they were present. In addition, in the presence of annoying ads, people were less accurate on questions pertaining to what they had read. None of these effects on users are desirable from the publisher's perspective, regardless of whether they are due to distraction or a lack of customer engagement.

Thus, when strategizing which ads to run, managers should consider not just the short-term revenue that the ads bring but the more subtle and long-term effects these ads may have on user retention and revenue.

REFERENCES

Aaker, David A. and Donald E. Bruzzone (1985), "Causes of Irritation in Advertising," *Journal of Marketing*, 49 (Spring), 47–57.

Bellman, Steven, Anika Schweda, and Duan Varan (2010), "The Residual Impact of Avoided Television Advertising," *Journal of Advertising*, 39 (1), 67–81.

Benway, Jan P. (1998), "Banner Blindness: The Irony of Attention Grabbing on the World Wide Web," *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, 42 (5), 463–67.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling (2011), "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?" *Perspectives on Psychological Science*, 6 (1), 3–5.

Burke, Moira, Anthony Hornof, Erik Nilsen, and Nicholas Gorman (2005), "High-Cost Banner Blindness: Ads Increase Perceived Workload, Hinder Visual Search, and Are Forgotten," *ACM Transactions on Computer-Human Interaction*, 12 (4), 423–45.

Buscher, Georg, Edward Cutrell, and Meredith Ringel Morris (2009), "What Do You See When You're Surfing? Using Eye Tracking to Predict Salient Regions of Web Pages," in *Proceedings of the ACM Conference on Human-Computer Interaction (SIGCHI '09)*. New York: Association for Computing Machinery, pp. 21–30, [DOI: 10.1145/1518701.1518705].

Carlson, Michelle C., Lynn Hasher, S. Lisa Connelly, and Rose T. Zacks (1995), "Aging, Distraction, and the Benefits of Predictable Location," *Psychology and Aging*, 10 (3), 427–36.

Chen, Mon-Chu, John R. Anderson, and Myeong-Ho Sohn (2001), "What Can a Mouse Cursor Tell Us More? Correlation of Eye/Mouse Movements on Web Browsing," in *Proceedings of the ACM Conference on Human-Computer Interaction (SIGCHI '01)*. New York: Association for Computing Machinery, pp. 281–82, [DOI: 10.1145/634067.634234].

Connelly, S. Lisa, Lynn Hasher, and Rose T. Zacks (1993), "Age and Reading: The Impact of Distraction," *Psychology and Aging*, 6 (4), 533–41.

Drèze, Xavier and François-Xavier Hussherr (2003), "Internet Advertising: Is Anybody Watching?" *Journal of Interactive Marketing*, 17 (4), 8–23.

Duchowski, Andrew T. (2007), *Eye Tracking Methodology: Theory and Practice*. London: Springer.

Edwards, Steven M., Hairong Li, and Joo-Hyun Lee (2002), "Forced Exposure and Psychological Reactance: Antecedents and Consequences of the Perceived Intrusiveness of Pop-Up Ads," *Journal of Advertising*, 31 (3), 83–95.

Gartner Inc. (2013), "Gartner Says Worldwide Mobile Advertising Revenue to Reach $11.4 Billion in 2013," press release, (January 17), (accessed September 17, 2014), [available at http://www.gartner.com/newsroom/id/2306215].

Gilbert, Daniel T., Douglas S. Krull, and Brett W. Pelham (1988), "Of Thoughts Unspoken: Social Inference and the Self-Regulation of Behavior," *Journal of Personality and Social Psychology*, 55 (5), 685–94.

Goldfarb, Avi and Catherine Tucker (2011), "Online Display Advertising: Targeting and Obtrusiveness," *Marketing Science*, 30 (3), 389–404.

Goldstein, Daniel G., R. Preston McAfee, and Siddharth Suri (2011), "The Effects of Exposure Time on Memory of Display Advertisements," in *Proceedings of the 12th ACM Conference on Electronic Commerce (EC '11)*. New York: Association for Computing Machinery, pp. 49–58, [DOI: 10.1145/1993574.1993584].

——, ——, and —— (2012), "Improving the Effectiveness of Time-Based Display Advertising," in *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. New York: Association for Computing Machinery, pp. 623–38, [DOI: 10.1145/2229012.2229058].

——, ——, and —— (2013), "The Cost of Annoying Ads," in *Proceedings of the 22nd International World Wide Web Conference (WWW '13)*, pp. 459–70, [available at http://www2013.org/proceedings/p459.pdf].

Guo, Qi and Eugene Agichtein (2010), "Towards Predicting Web Searcher Gaze Position from Mouse Movements," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. New York: Association for Computing Machinery pp. 3601–06, [DOI: 10.1145/1753846.1754025].

Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011), "The Online Laboratory: Conducting Experiments in a Real Labor Market," *Experimental Economics*, 14 (3), 399–425.

Huang, Jeff, Ryen W. White, and Georg Buscher (2012), "User See, User Point: Gaze and Cursor Alignment in Web Search," in *Proceedings of the 2012 Annual Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, pp. 1341–50, [DOI: 10.1145/2207676.2208591].

IAB (2013), "IAB Internet Advertising Revenue Report: 2012 Full Year Results," (April), (accessed September 17, 2014), [available at http://www.iab.net/media/file/IABInternetAdvertisingRevenueReportFY2012POSTED.pdf].

Li, Hairong, Steven M. Edwards, and Joo-Hyun Lee (2002), "Measuring the Intrusiveness of Advertisements Scale Development and Validation," *Journal of Advertising*, 31 (2), 37–47.

Mason, Winter and Siddharth Suri (2012). "Conducting Behavioral Research on Amazon's Mechanical Turk," *Behavior Research Methods*, 44 (1), 1–23.

McCoy, Scott, Andrea Everard, and Eleanor T. Loiacono (2008), "Online Ads in Familiar and Unfamiliar Sites: Effects on Perceived Website Quality and Intention to Reuse," *Information Systems Journal*, 19 (4), 437–58.

Mueller, Florian and Andrea Lockerd (2001), "Cheese: Tracking Mouse Movement Activity on Websites, a Tool for User Modeling," in *CHI'01 Extended Abstracts on Human Factors in Computing Systems*. New York: Association for Computing Machinery, pp. 279–80.

Navalpakkam, Vidhya and Elizabeth Churchill (2012), "Mouse Tracking: Measuring and Predicting Users' Experience of Web-Based Content," in *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, pp. 2963–72, [DOI: 10.1145/2207676.2208705].

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis (2010), "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, 5 (5), 411–19.

Parducci, Allen and Linda F. Perrett (1971), "Category Rating Scales: Effects of Relative Spacing and Frequency of Stimulus Values," *Journal of Experimental Psychology*, 89 (2), 427–52.

Riley, John G. (2001), "Silver Signals: Twenty-Five Years of Screening and Signaling," *Journal of Economic Literature*, 39 (2), 432–78.

Smit, Annika S., Paul A.T.M. Eling, and Anton M.L. Coenen (2004), "Mental Effort Causes Vigilance Decrease due to Resource Depletion," *Acta Psychologica*, 115 (1), 35–42.

Tavassoli, Nader T. (2008), "The Effect of Selecting and Ignoring on Liking," in *Visual Marketing: From Attention to Action*, Michel Wedel and Rik Pieters, eds. New York: Lawrence Erlbaum Associates, 73–89.

Toomim, Michael, Travis Kriplean, Claus Pörtner, and James A. Landay (2011), "Utility of Human–Computer Interactions: Toward a Science of Preference Measurement," in *Proceedings of CHI 2011: ACM Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, pp. 2275–84, [DOI: 10.1145/1978942.1979277].

Venables, W.N. and B.D. Ripley (2002), *Modern Applied Statistics with S*. New York: Springer.

Willemsen, Martijn C. and Eric J. Johnson (2010), "Visiting the Decision Factory: Observing Cognition with MouselabWEB and Other Information Acquisition Methods," in *A Handbook of Process Tracing Methods for Decision Making*, M. Schulte-Mecklenbeck, A. Kühberger, and R. Ranyard, eds. New York: Taylor & Francis, 21–42.

Yoo, Chan Yun and Kihan Kim (2005), "Processing of Animation in Online Banner Advertising: The Roles of Cognitive and Emotional Responses," *Journal of Interactive Marketing*, 19 (4), 18–34.