# The Economic and Cognitive Costs of Annoying Display Advertisements

Daniel G. Goldstein
Principal Researcher
Microsoft Research, New York City

Siddharth Suri
Senior Researcher
Microsoft Research, New York City

R. Preston McAfee
Chief Economist
Microsoft

Matthew Ekstrand-Abueg
Graduate Student
Northeastern University

Fernando Diaz
Senior Researcher
Microsoft Research, New York City

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, & Siddharth Suri. (2013). The cost of annoying ads. Proceedings of the 23rd International World Wide Web Conference. 

ACKNOWLEDGEMENT
We thank Randall A. Lewis, Justin M. Rao, and David H. Reiley for helpful conversations.

# The Economic and Cognitive Costs of Annoying Display Advertisements


## *Abstract*

Some online display advertisements are annoying. While publishers know the payment they receive to run annoying ads, little is known about the cost such ads incur, for instance, by causing website abandonment. Across three empirical studies, we address two primary questions. What is the economic cost of annoying ads to publishers? What is the cognitive impact of annoying ads? First we conduct a preliminary study to identify sets of more and less annoying ads. Second, in a field experiment, we calculate the compensating differential, the amount of money one would need to pay users to generate the same number of impressions in the presence of annoying ads as they would generate in their absence. Third, we conduct a mouse-tracking study to investigate how annoying ads may affect reading processes. We 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

In the online display advertising industry, advertisers pay publishers (websites) to run display ads that are seen alongside other content by users (website visitors). 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 had revenues of over $12 billion in the United States, or 33% of total online advertising revenue (IAB, 2012, p. 12). Online advertising itself brings in about as much revenue as broadcast television, and more revenue than cable television, radio, and newspaper advertising (IAB, 2012, 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 over 200 million downloads of one ad blocker alone[1]. From an economic perspective, 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 thus 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 who are concerned with maximizing short-run sales commissions and parties concerned with maximizing long-run user engagement. Anecdotally, we have heard this described as the "religious war" over annoying ads. The presence of annoying ads might also signal that a publisher is desperate for

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

business. In the case of a publisher that provides vital services, such as email, such apparent desperation might cause users to switch to providers that seem to be flush with resources and therefore 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 virus, spyware, and malware infections. And, of course, being annoyed is a cost in itself.

For advertisers, annoying ads gain user attention, but the downsides may be several. The use of annoying ads may cause users to distance themselves from an advertiser's brand or to question its reputability (McCoy et al. 2008). And, as was said of publishers, advertisers who choose the route of annoyance may 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 highly annoying ads may be less likely to be remembered by users (Yoo and Kim, 2005) and that actively-ignored stimuli like 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, they may increase the use of ad-blocking software (Edwards et al 2002, Li et al 2002). At the limit, the use of ad blockers could reduce the number of publishers, leaving advertisers with fewer places to advertise.

We intend to shed light on two main questions concerning annoying display ads:

*What is the economic cost of annoying ads?* Annoying ads presumably have a cost to publishers arising from user abandonment, but 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 one would need to pay users to generate the same number of impressions in the presence of annoying ads as they would generate in their absence.

*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.

In what follows, 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. Experiment 2 investigates 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 implications of this research for managers.

To begin, we review the relevant literature. A number of marketing investigations have explored causes of annoyance in television advertising (Aaker & Bruzzone, 1985, Bellman et al 2010). Our focus is internet advertising, which has received less attention. Drèze and Hussherr (2003), used eye-tracking technology on participants viewing web pages. They found 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 found that the presence of ads increased the time it takes people to conduct visual searches, with no significant difference between animated and static ads. (We note that animated and static is not necessarily the same as

annoying and not annoying, and will look at the relationship in the preparatory study.) They also found that animated ads were 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 found that moderate animation had a positive effect on advertisement recognition rates as well as on brand attitude measures. They also found that rapidly animated (presumably annoying) banner ads can backfire, leading to lower recognition rates and more negative attitudes towards the advertiser. This finding, combined with the work of Burke et al (2005), lends support to the idea that annoying ads can have negative effects not only on users and publishers, but also on advertisers. In a field experiment, Goldfarb and Tucker (2011) identified two types of ads that increase self-reported purchase intentions: those that were intrusive or those that matched a site's content. They found, however, ads that had both properties, reduced purchase intentions. In sum, 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 the methodology of Toomim et al. (2011), in which experimental participants are randomly assigned tasks of varying difficulty for randomly-assigned rates of pay. For instance, in one study, Toomim and colleagues pay people to transcribe images of text on webpages that are randomly assigned to be user-friendly or user-unfriendly. Pay rates are also randomly assigned. By analyzing the amount of work done in each condition, the authors could compute the compensating differential, that is, the amount of additional money one would have 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 shall use a similar method to estimate the effect of ad quality on Web site abandonment.

## *Preparatory Study: Qualitative assessment of ad annoyance*

The preparatory study has several objectives. The first is to generate sets of annoying and non-annoying (henceforth "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 the Amazon Mechanical Turk online labor market (Paolacci, Chandler & Ipeirotis, 2010; Horton, Rand & Zeckhauser, 2011; Buhrmester, Kwang & Gosling, 2011; Mason & Suri 2012) on which participants were paid 25 cents plus a bonus of 2 cents per ad rated. The task was restricted to US-based participants who had at least a 95% approval rating. From 163 US-based participants who began the task, we analyze the 141 participants who skipped at most one of the 36 questions.

At the start of the experiment, in order to familiarize them with the range of stimuli (Parducci, 1971), participants were shown 36 ads (4 ads per page over 9 pages), but were not asked to rate them. 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 by 250 pixels) and 36 of which were skyscrapers (120 or 160 by 600 pixels). From each of these 72 animated ads, a static variant was created 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

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

that the causal impact of animation on annoyance could later be measured. 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 5 point scale ranging from: "Much less annoying than the average ad in this experiment" to "Much more annoying than the average ad in this experiment". Categories were chosen 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 they rated, but never the animated and static variant of the same original ad. After this rating task, each ad that a participant rated as annoying was presented 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 1-5 scale. Because participants were told that category 3 was to represent "average" annoyingness for this experiment, the 2.9 rating reflects good aggregate calibration.

[INSERT FIGURE 1 ABOUT HERE]

FIGURE *1* plots the mean annoyingness rating of the 72 ad pairs. One striking result is that animated ads were rated as a great deal 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, while on the other end, the 24 least annoying ads were all static. The 10

most and least annoying ads (per the mean ratings) are designated the “bad” and “good” ad sets for use in Experiments 2 and 3. Note that this implies that the bad ads all turned out to be animated and the good ads all turned out to be static.

Participants who rated an ad as annoying (categories 4 or 5) were asked to type reasons why the ad was found to be annoying. They submitted 1846 textual responses of this type. Based on 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 less than 10 times and “stop words” resulted in a short list of common, substantive words. We categorized each substantive word into one of the 5 relevant categories where possible. We then went back to the original long list, assigned each word of participant input to one of the five categories where possible, and tabulated the counts.

The most common reason given for an ad being annoying, by a fair margin, was animation. The “animation” category (typified by words like “move”, “motion” and “animate”) occurred 771 times in the sample. 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” or “distracting”), rather than the ad features 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 and Bruzzone’s (1985) list of characteristics that make television ads annoying. Next most often (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 a category of Aaker and Bruzzone (1985) in which “the situation is contrived, phony, unbelievable, and/or overdramatized”.

This preparatory study achieves two goals. It provides us with sets of more and less annoying ads for Experiments 1 and 2, and finds 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 may 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 et al 2002, Li et al 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 non-annoying ad sets based on user ratings, and not on ad features. Our primary focus is understanding 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 impact intent to return to a website (McCoy et al 2008, Li et al 2002). 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 the method of Toomim et al (2011). Experimentation was carried out on Amazon’s Mechanical Turk labor market and participants were 1223 workers with approval ratings or 90% or more. Payment was advertised as a flat rate of 25 cents plus a bonus. Since the bonus was randomly assigned, it was not revealed until the task was accepted in order to prevent

self-selection on the basis of bonus pay. The task was advertised as involving email classification, which is a common kind of job in the Mechanical Turk labor market. Upon accepting the task, workers were told that they would be shown emails and asked to classify them as "Spam", "Personal", "Work", or "e-Commerce" related. We chose this task because of its realism and because it has the property that people could quit (i.e., stop categorizing emails) at any time. The number of emails categorized was a revealed choice and our primary dependent measure. The emails used in the experiment were randomly drawn from the public-domain Enron dataset[3], which provides ground-truth data as to whether each email is spam or not. The ground truth data allowed us to test whether email classification accuracy would depend on pay rate and annoyingness of advertising.

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 emails classified, of one, two, or three cents. 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 10 least and 10 most annoying ads as determined in the preparatory study. No mention was made of advertising or randomized pay conditions. The exact bonus amount was only revealed to participants after they agreed to undertake the task. We ran a chi-squared test to check for significant differences in the number of participants choosing to begin the task across the 9 conditions and found none (Chi square test, $p = 0.25$).

In the experiment, participants were shown one email per page with either 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

[3] http://www.cs.cmu.edu/~enron/. We removed phone numbers, email addresses and the word "Enron" from the emails in order to safeguard privacy and to reduce distraction.

and right of the text were randomly selected, at each page load, from the relevant set of 10 good or 10 bad ads from the preparatory study. At the bottom of each page were radio buttons to allow the participant to classify the email into the four categories and buttons to either classify another email or to quit the experiment. The text width and page width were such that the page would be visible without scrolling for the vast 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.

[INSERT FIGURE 2 ABOUT HERE]

Each email to be classified was shown on a new page, so viewing one email constituted one impression. As soon as the task was accepted, one email was presented, meaning that each participant generated at least one impression. The primary dependent measure is the number of impressions (i.e., emails classified) per person per condition. The mean number of emails classified was 61. The median was 25 and the first and third quartiles were 6 and 57, reflecting strong skewness. Only two of the 1223 participants reached the upper limit of classifying 1000 emails. Means and standard errors per condition are given in Table 1. The randomly assigned ad quality did not affect the likelihood of classifying an email as spam, which was similar across conditions (47.5%, 50% and 48.5% in the bad, good, and none conditions respectively; p=.975 by Chi square test).

[INSERT TABLE 1 ABOUT HERE]

[INSERT FIGURE 3 ABOUT HERE]

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 1 cent pay rate, the no ad condition is lower than the ad conditions. However, at the 1 cent pay rate, there is no significant difference in the number of emails classified according to ad condition, either by an ANOVA ($p=.42$) or by inspecting 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 & 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. Here it can be seen that good ads and no ads are predicted to have a similar effect, with bad ads causing substantially fewer impressions.

[INSERT TABLE 2 ABOUT HERE]

[INSERT FIGURE 4 ABOUT HERE]

Given the non-linear 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. Therefore, the pay raise required to match the effect of moving from bad ads to no ads is .153 cents per impression (.2 *

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

12.68/16.58). In other words, a participant in the bad ads 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. Or, in CPM (cost per thousand impression) terms, the cost of bad ads in this experiment was $1.53 per thousand impressions. Interestingly, many "bad ads" pay less than $1.53 CPM. In fact, recently, 53% of all display ad impressions were estimated to pay between 10 and 80 cents 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 emails 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 we calculate here 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 ads condition would need to be paid $1.15 per thousand impressions to generate as many impressions as a participant in the good ads condition. Again, if these estimates generalize, this suggests that bad ads could lose money since they may only pay publishers $0.50 CPM or less. By a similar calculation, we see that the cost of good ads relative to no ads is $0.38 CPM, noting that this estimate is based on a statistically insignificant difference.

To this point, we have looked at the effect of annoying ads on dropout, which is a main concern of publishers. We turn now to the users' perspective and look at the effects of annoying ads on a cognitive task. Recall that for each email classified, the Enron email corpus contained ground

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

truth information as to whether it was spam or not spam, which allows us to test the effect of ad types on email classification accuracy. In general, classification accuracy was rather high at 91%. Table 3 shows the results of two regressions predicting individual accuracy rates, controlling for the number of emails categorized. Against a baseline of annoying ads, people classified emails more accurately in the presence of good ads or no ads. Because ad conditions were randomly assigned, it appears that annoying ads have a causal impact on accuracy. The regressions imply that accuracy drops about 1 percentage point per 128 emails 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 group of people that both persists in the presence of bad ads and does not care about accuracy. To test for this, we looked at accuracy binned by the number of emails categorized. Across all four quartiles of the distribution of emails categorized, the accuracy of the people assigned to the bad ads condition was about 2 to 3 percentage points lower (3.1%, 2.9%, 1.9%, 2.0% from lowest to highest quartile of emails categorized) than those in the no ads condition. Similarly, a regression does not find a significant interaction between the bad ad condition and the number of impressions. This suggests that the negative impact of bad ads on accuracy is stable relative to the number of emails classified and not due to self-selected dropout.

[INSERT TABLE 3 ABOUT HERE]

We have seen thus far that annoying ads cause people to abandon paying tasks, and that they seem to have a negative effect on a cognitive task, classifying emails. What we do not yet know is why. For instance, annoying ads could have affected accuracy because they distracted people and impaired the reading process, or because they signaled to people that the site creators do not

care about them, causing workers 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, bad ads seem to cause participants to drop out earlier and to exhibit lower classification accuracy. The motivation behind Experiment 2 is to gain some insight about why these effects arise.

One way to understand psychological processes, especially those involving tasks done at a computer, is by eye tracking (Buscher et al, 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 hard 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, & Sohn, 2001; Guo & Agichtein, 2010; Huang, White & Buscher, 2012). It is also thought to proxy for user interest (Willemsen & Johnson, 2010; Navalpakkam, & Churchill, 2012). Because the necessary Javascript code can be embedded into a webpage, mouse-tracking kinds of studies can be done cheaply and at scale online (Mueller & 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 the *distraction hypothesis*: the dropout and lower

accuracy are due to annoying ads disrupting the reading process. If distraction is at play, we would expect people 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 in empirical observations that distraction increases reading time (e.g., Connelly, Hasher & Zacks, 1991; Carlson, Hasher, Connelly & Zacks, 1995). A few reviewers of this felt that distraction might increase reading speed or cause less deliberate reading. To accommodate this, we will 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 good ads in the experiment, but reading behavior on 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 will speculate a bit about what those might be. Though it does not pin down an exact mechanism behind dropout and lower accuracy, this experiment can also be seen as a qualitative investigation of what annoying ads do to the experience of consuming web content.

Experimentation was conducted on Amazon's Mechanical Turk. We restricted our participant pool to those living in the United States and 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 cent (US) flat rate plus 10 cents per question answered. After a text passage was read, four questions were asked. We paid per question answered as opposed to

paying for correct answers to remove an incentive for participants to share answers outside of 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 the layout of the page to be uniform across all browsers and screen sizes. Rendering the article as an image also ensured the URLs in the web page could not be followed, however, attempted clicks were recorded. The text of the article consisted of a story regarding school teachers and had an accompanying graphic (see Figure 5, left panel). Participants were randomly placed into one of 3 treatments: bad ads, good ads or no ads. In the bad and good ad conditions a randomly chosen bad or good ad was placed 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 whitespace geometry and making it only suitable for measuring page-viewing time and comprehension, but not for mousetracking (due to "parking" effects, as will be explained).

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 used exclusively skyscraper-dimension ads to ensure the page layout would be identical in the good and bad ad conditions so they can be compared directly—when page content and page layout both change in a mouse-tracking study, one cannot 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 5 or 6 possible answers. The answer to the last of the three questions resided in the second to last sentence of the article. This tested if a participant read to the end of the article or not. 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. Participants were randomly assigned to conditions and there were no significant difference in cell counts (Chi square, $p = .54$). As a manipulation check to make sure that the annoying ads condition was indeed annoying, we coded responses to the fourth question as to whether it referred to being annoyed by an ad or not. 41.5% of participants complained in the bad ad condition, 4.5% complained in the good ad condition and 0% complained in the no ad condition. The annoying ads seem to have annoyed people.

We begin by first checking if the ad treatment had an effect on the amount of time participants spent on the page. Participants in the bad ads condition spent on average 73.1 seconds on the page with a standard error of 1.2 seconds, while participants in the good ads 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 ads treatment had a mean of 71.9 seconds and a standard error of 1.4 seconds. Furthermore, participants in the bad ads condition took longer to view the page than those in the combined “innocuous” (good and no ad) conditions. ($p=.02$, ANOVA of log time data). The difference to the good condition alone was significant ($p=.02$), while the difference to 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) as it may involve looking at the ad as well as at the text. We shall use the mousetracking data to get a proxy for how much time was spent attending to the ad and text separately.

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

[INSERT FIGURE 5 ABOUT HERE]

Visual inspection of heatmaps 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 heatmaps are more exploratory, in what follows we will quantitatively test fixations and other measures of reading behavior.

[INSERT TABLE 4 ABOUT HERE]

To understand of how reading behavior may change according to condition, we measure and report on a number of dependent variables with the mousetracking 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 below shows all of these dependent variables for both ad treatments. A consistent story emerges. 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, compared to good ads. We tested the significance of these effects by log transforming them and comparing the means between these two treatments. ANOVAs were run to test significance and all of the effects were significant, as seen in Table 4. Notably, people made 183% more ad fixations and spent 70% more time on the bad ads than the good ones. When modeling mouse movements on the basis of treatment, we find that the 41.5% who complained about the ads in the bad ads condition showed

even stronger effects when compared to the good ads condition, but withhold the results for brevity.

Using the mouse as a proxy for attention, it seems that annoying ads get noticed more than benign ads. This alone might explain why people 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 question next.

Focusing our analysis on text (as opposed to the ad) of the article, we measure the same dependent variables. As shown in the bottom rows of Table 4, and perhaps surprisingly given the ad results, there is not an 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, multiple comparisons, and 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 onto the next page) and obtain the same basic results. As a second robustness check, we compared the 41.5% percent who were annoyed by the ad in the bad ad condition (according to self-reports) to the good ad condition. A similar pattern of non-significant differences appeared, and even the fixations measure became non-significant (p=.20), despite there being a large number of observations (959 in the good ads condition, 399 in the bad ads condition) in the regression.

As an additional check, we wanted to assure ourselves 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 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 whitespace to the immediate right of any text or graphics. It is for this reason that we cannot directly compare the "no ad" condition with the good and bad ad conditions for mousetracking. 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 ads condition, 54.2% of participants parked the mouse in the text area, compared to 56.2% in the good ads 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. It appears as if the results in Table 4 are not due to parking.

In contrast to the prediction of a distraction hypothesis, it seems as if 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, right panel, 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 will be seen, about five percent of participants exhibited this kind mouse reading behavior. Under the distraction hypothesis we would expect to see 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

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

with the results shown in Table 4—then the likelihood of mouse reading would be unaffected. To test this, three judges, blind to the conditions, rated all 1,977 mouse maps from the bad and good ad conditions and rated each as to whether it was an instance of mouse reading. Pairwise agreement between the three raters was 94.3%, 95.4% and 96.4%. Using a majority criterion to classify maps, mouse reading was observed in 67 out of 989 (6.8%) participants in the bad ads condition, and 71 out of 988 participants in the good ads condition (7.2%), an insignificant difference (p=.786 by Chi square test). Using a unanimity criterion gave a very similar result (4.8% vs 5.9%, p=.313 by Chi square test). Thus far, the mousetracking analysis suggests that the bad ads receive more attention than the good ads, but that reading behavior was not affected by the annoyingness of ads.

Recall that we asked three reading comprehension questions. We defined an overall accuracy index, ranging between 0 and 3, which is simply the number of questions a participant answered correctly. With this measure we see the bad ads condition as significantly less accurate than the no ad condition using the Tukey Honest 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 by ANOVA). Two of the three questions were apparently quite easy (greater than 95% correct in all treatments) and had a ceiling effect in their results. The third question, which occurred at the very 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 none conditions, respectively). A reviewer of this work suggested that mouse tracking may reveal whether people pay less attention to the last paragraph when bad ads are present. To test this, we defined a subset of the text rectangle that just includes

the last four lines—lines which held the answer to the third question—and measured the same metrics as in Table 4. The result is mixed. As one might expect, people in the bad ads condition (who were more likely to get the question wrong) had fewer entrances into the relevant area than people in the good ads condition (M=1.95, SD=.06 vs M=2.18, SD=.08, p=.004 by 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=8049, SD=588 vs. M=7748, SD=526, p=.804). Perhaps because only 5% of people read line by line with the mouse (as noted above), mousetracking may be too blunt of an instrument to detect attention to very small areas of a page, and such a test would be better suited for eye-tracking.

Taking these results as a whole, if mouse movements are good proxies for attention on large sections of a page, users' attention does seem to be captured by annoying ads. In the ad space, time, distance, fixations and entrances were all significantly greater for bad ads than good ads. Furthermore, these results to not seem to be due to artifacts, such as users trying to click the ads or users 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. Interestingly, annoying ads did not seem to affect mouse behavior reading behavior or time spent on the text area. To gain some more insight into the reading process, we looked at 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 loads, and then move it back towards the text. This 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 mousetracking results in this paper and with recent investigations of online ad exposure (Goldstein, McAfee & Suri, 2011; Goldstein, McAfee & Suri, 2012), which suggest that the whole page is scanned just after it loads followed by a period in which people 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 could be the case that working to ignore ads over an extended time is somehow cognitively depleting (Smit, Eling & Coenen, 2004; Gilbert, Krull & Pelham, 1988) and leaves users with fewer resources to be accurate. To investigate this idea, we looked at the data from Experiment 1 to observe whether the deleterious effect of bad ads on accuracy increased over time. However, in contrast to this idea, we found the difference to the "no ads" condition to be rather constant. An alternative account as to why bad ads harmed accuracy is that users expressed their dissatisfaction with the annoying ads by exerting lower effort on the email classification and reading comprehension questions. To get a satisfactory answer as to why annoying ads cause dropout and decreased accuracy, further research is needed.

## *Conclusions*

Summing up the empirical work in this paper, the preparatory study found that some ads are perceived as much more annoying than others. Among the complaints, animation was pre-eminent and exerted a causal effect on annoyingness ratings. Poor aesthetics and questionable advertiser reputability were also frequently mentioned. Experiment 1 showed 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 email classification task. This experiment allowed us to estimate the compensating differential. In our study, to motivate an individual to generate as many impressions in the presence of bad ads as they would in the presence of no ads or good ads, one needs 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. It found that annoying ads garnered significantly more attention than controls, as proxied by the mousetracking 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 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 does indeed lose money depends on many factors, including the alternatives in the market. In our studies, the participants' alternative was finding another task on Mechanical Turk or finding something else to do on the internet altogether. If annoying ads are running on a very unique and valuable site (e.g., imagine there were only one free email provider in the world), then users' tolerance for annoying ads would be expected to be very high. On the other hand, if annoying ads are running on a site that offers what many other sites do (e.g., news stories from mass-market newswires), then switching costs are low and people's tolerance for annoying ads would be estimated as 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 react, perhaps by charging advertisers for the externalities they impose.

We conclude with the two main questions that motivated this work:

*What is the economic cost of annoying ads to publishers*? We see in a field study 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, in order to retain users. This short-term cost estimate can be thought of as a lower bound on the total cost of annoying ads. There may be longer term costs. For instance, upon dropping out, users may decide never to return to a site that annoyed them.

*What is the cognitive impact of annoying ads*? In our studies, people seem to notice annoying ads, 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 concerning what they have 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.

When considering what ads to run, it makes sense for managers to consider not just the short-run revenue that ads bring, but the subtler long-run effects they may have on user retention and revenue.

# References

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

Bellman, Steven, Anika Schweda, 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". In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, Vol. 1, pages 463–467.

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*, 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):423–445.

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)*.

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),* 427-436.

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)*.

Connelly, S. Lisa, Lynn Hasher, and Rose T. Zacks. (1993) "Age and reading: the impact of distraction." *Psychology and Aging 6(4),* 533-541.

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

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

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," 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, 685-694.

Goldfarb, Avi and Catherine Tucker (2011), "Online display advertising: Targeting and obtrusiveness," *Marketing Science*, 30(3):389-404.

Goldstein, Daniel G., R. Preston McAfee, & Siddharth Suri. (2013). The cost of annoying ads. *Proceedings of the 22nd International World Wide Web Conference (WWW '13*), 459-470.

Goldstein, Daniel G., R. Preston McAfee, and Siddharth Suri (2012), "Improving the Effectiveness of Time-Based Display Advertising," In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*.

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)*.

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*, pp. 3601-3606. ACM.

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*.

IAB 2012. IAB internet advertising revenue report: 2012 full year results. 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, 437-458.

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*, pp. 279-280. New York: Association for Computing Machinery.

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*, pp. 2963-2972. New York: Association for Computing Machinery.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis (2010), “Running Experiments on Amazon Mechanical Turk,” *Judgment and Decision Making*, 5, 411-419.

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), 427-452.

Riley, John G. (2001), “Silver Signals: Twenty-Five Years of Screening and Signaling,” *Journal of Economic Literature*, XXXIX, 432-478.

Smit, Annika S., Paul A. T. M. Eling, 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: NY, 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*.

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

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, 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.

# *Tables*

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) |

*Note: One impression is one email classified. Standard errors are in parentheses.*

TABLE 2:

MODELS PREDICTING NUMBER OF IMPRESSIONS BASED ON THE TYPE OF AD PRESENT.

| | Model 1 | Model 2 |
|---|---|---|
| Intercept | 3.43 (.12)*** | 3.43 (.12) *** |
| Good ads | 0.17 (0.10)+ | |
| No ads | .22 (0.10)* | |
| Good or No ads | | 0.19 (0.08) * |
| Pay rate | 26.47 (4.8)*** | 26.61 (4.8) *** |
| | | |
| AIC | 12158.57 | 12156.85 |
| Nagelkerke Pseudo $R^2$ | .04 | .04 |
| Log Likelihood | -6074.29 | -6074.43 |
| Deviance | 1481.00 | 1481.04 |
| Number of observations | 1223 | 1223 |

*** p<0.001,** p<0.01,* p<0.05, + p<.1

*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 given in parentheses. In model 2, good ads and no ads are treated as one category.*

TABLE 3:

MODELS PREDICTING CLASSIFICATION ACCURACY RATE

| | Model 3 | Model 4 |
|---|---|---|
| Intercept | 0.90 (0.01)*** | 0.90 (0.01)*** |
| Impressions | -7.8e-5 (0.00)** | -7.8e-5 (0.00)** |
| Good Ads | 0.02 (0.01)* | |
| No ads | 0.03 (0.01)** | |
| Pay rate | 0.14 (0.43) | 0.14 (0.43) |
| Good ads or no ads | | 0.02 (0.01)** |
| $R^2$ | .02 | .01 |
| Number of observations | 1057 | 1057 |

*** p<0.001,** p<0.01,* p<0.05

*Notes: Impressions refers to the number of emails categorized. Models exclude people who did not classify any emails, as 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 given in parentheses.*

TABLE 4:

MOUSE-TRACKING METRICS

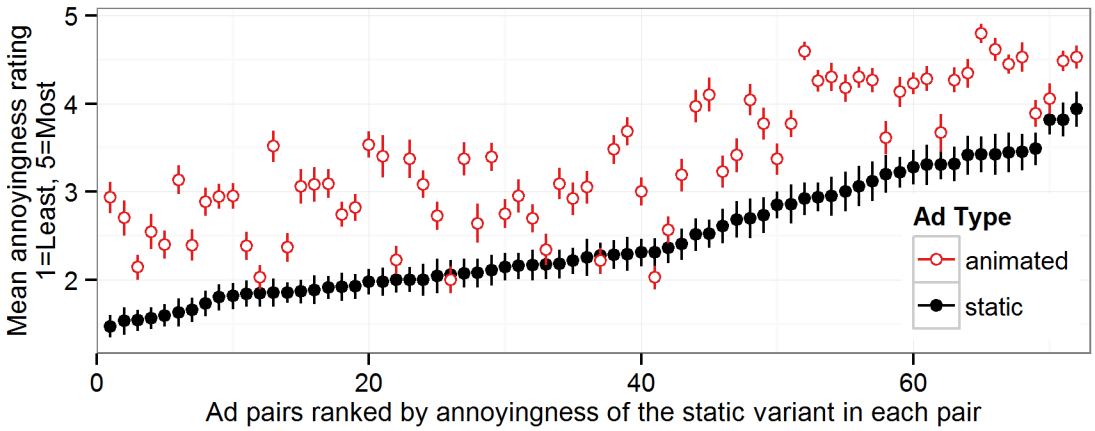
| Area | Measure | Bad Ads | Good Ads | P-Val (log OLS) |
|---|---|---|---|---|
| Ad | Fixations | 4.45 (0.67) | 1.57 (0.26) | <.001 |
| | Distance | 182.6 (7.9) | 157.9 (8.0) | .003 |
| | Entrances | 1.31 (0.05) | 1.13 (0.05) | .004 |
| | Time (ms) | 1873 (321) | 1101 (186) | .001 |
| Text | Fixations | 135.7 (9.07) | 128.0 (9.05) | .047 |
| | Distance | 1492 (55.2) | 1570 (66.6) | .677 |
| | Entrances | 1.51 (0.06) | 1.50 (0.07) | .322 |
| | Time (ms) | 38268 (1207) | 36637 (1085) | .596 |

*Notes: Top rows: Mouse-tracking metrics from the advertisement area (rectangle to the right of the text). All of 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 seem to change substantially in the text area as a function of ad condition*

# *Figures*

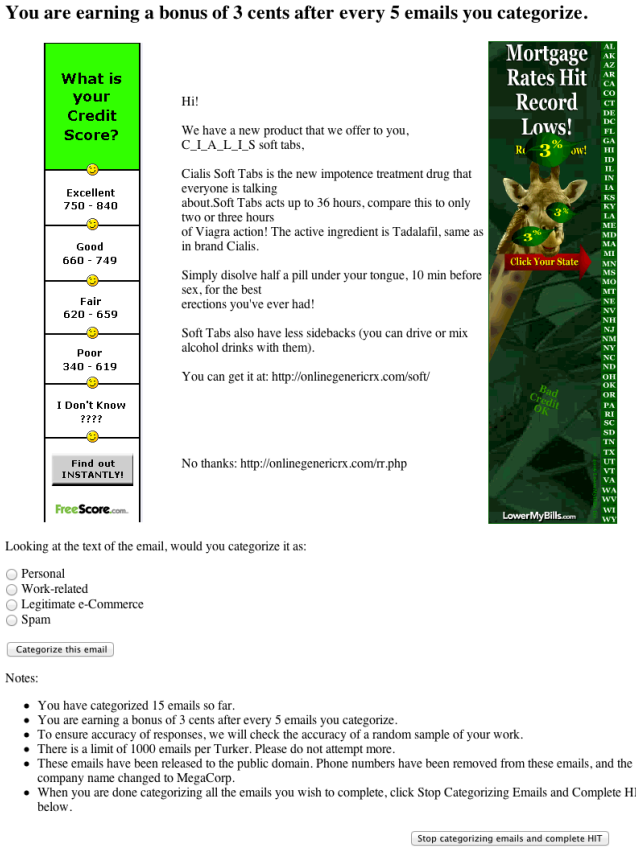
FIGURE 1:

MEAN ANNOYINGNESS RATINGS OF ADS




*Note: 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, showing how animated ads were rated as much more annoying than animated ones. Error bars extend one standard error above and below the means.*

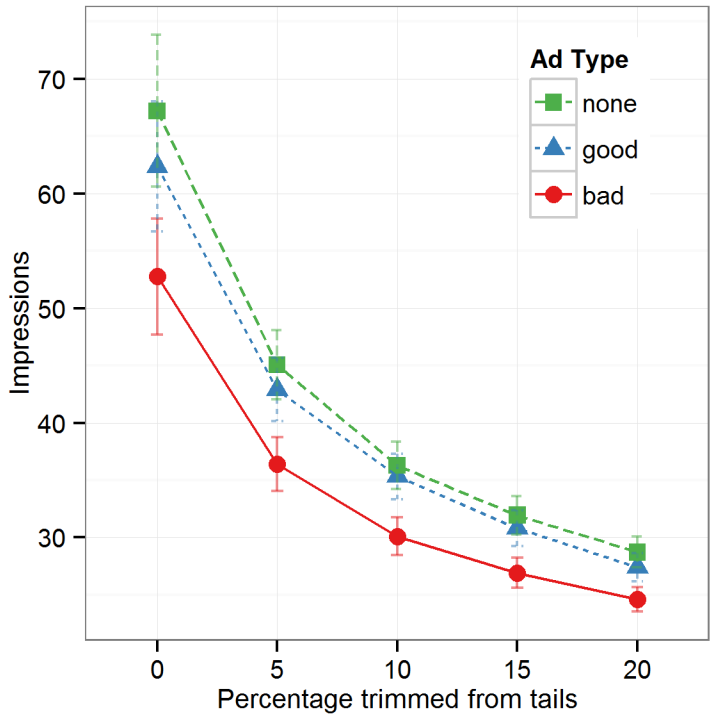
FIGURE 2: EMAIL CLASSIFICATION PAGE




*Note: Figures is from the "bad ad" condition with a pay rate of three cents per five emails classified. Radio buttons provide categories into which to categorize the email. Buttons allow participants to quit the task or classify another email. 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 emails categorized so far.*

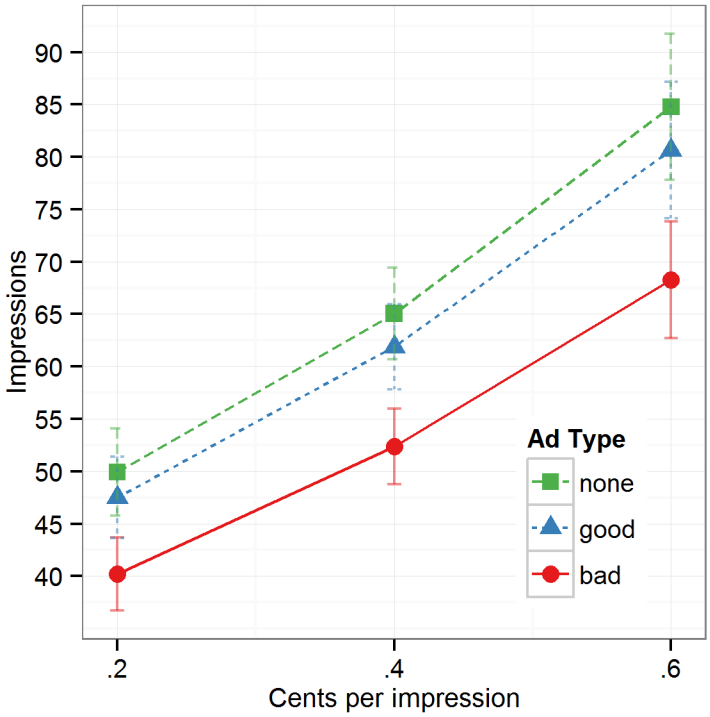
FIGURE 3:

ROBUSTNESS CHECK




*Note: 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.*

FIGURE 4: IMPRESSIONS BY PAY RATE AND AD CONDITION




*Note: Impressions refers to emails categorized. Error bars extend one standard error above and below the predicted values.*

FIGURE 5: HEAT MAP AND MOUSE MAP


*Notes: Left panel: Heat map of fixations for the bad ad condition in which red colors reflect more fixations and blue colors reflect fewer ones. Right panel: "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 (seen 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 in the lower left corresponds to the position of the "next" button.*