Effects of ad quality & content-relevance on perceived content quality

Henriette Cramer
Yahoo Labs
Sunnyvale, CA
henriette@yahoo-inc.com

ABSTRACT

Native advertising, ads that are highly cohesive with actual content in format and style, is a pervasive online trend. We report on a two-stage Mechanical Turk experiment exploring the effects of perceived quality of native ads on perceptions of site quality. First, a set of native ads was rated (N=98) for perceived annoyance and trust. Second, we compare four conditions of an aggregation of news headlines (N=237): 'no ads', 'low quality ads', 'high quality ads + content-relevant', 'high-quality ads + not content-relevant'. Our results indicate that native ads, which in isolation have been rated as *high quality*, could still have a *negative effect* on perceived site credibility and perceived site quality if they are too *content-relevant*. In addition to the effect of ad quality alone (e.g. non-annoying, trustworthy ads vs. intrusive scam ads), there is an additional impact for ads that are too similar to content; avoiding confusion is important for quality perceptions.

Author Keywords

Ads; advertising; content relevance; quality

INTRODUCTION

Online advertising is a pervasive and important aspect of online experiences, with advertising revenue supporting many major services. It is extremely beneficial for service and site designers to know what the effect of advertising formats will be on user perceptions and experience as a whole, especially considering potential negative user reactions to online advertising [1,4]. A current trend is native advertising, in which ads are presented in a manner that is highly cohesive with actual content in both format and style [9]. These ads are intended to overcome ad blindness, where users do not notice ads anymore [11]. They can be distinguished from 'display advertising' such as more traditional banners or right-rail ads. Native ad

examples include brands' 'suggested posts' on Facebook and Instagram, search result ads, and promoted tweets on Twitter. Industry studies such as [10] show that native ads are much more effective than display advertising in terms of user engagement and click-through rates. However, since these are such an integral part of the user experience, it becomes even more important to consider these ads' effects on the perceptions of services as a whole. Nevertheless, there are few such published studies.

The study presented here shows that native ads affect overall perceptions of websites in the context of news headline aggregations. We show that focusing on ad quality alone, without considering ad context will not provide sufficient insight in the effects on perceived site quality; some high quality ads may have unintended adverse effects.

BACKGROUND

Annoying ads have serious costs for both users and services; they impede not only users' task performance, but can also cause site abandonment [4]. Users often avoid online advertising, either through ad blindness, or actively (e.g., via ad blockers). Perceived goal impediment (ads getting in the way), previous negative experiences and perceived ad clutter are the main predictors of such avoidance [3]. Native ads are meant to overcome ad blindness and be more effective by being an integral part of the user's experience [10], fitting in with existing content format and structure.

In the context of search ads, the amount of attention paid to ads depends on whether they are relevant to the users' searches [2]. De Sa et al., [4] found that content relevance increased ad effectiveness in terms of ad recall, but did not lead to a better user experience. In contrast, personal relevance had little or no impact on recall, but did lead to a more pleasant user experience. For non-native display advertising, Goldfarb and Tucker [6] find that matching ads to website content, or making ads more visually obtrusive can lead to increased purchase intent. However, ads that combine both strategies (i.e., more visually apparent + content-relevant) decrease ad effectiveness. This may potentially be explained by increased ad intrusiveness also heightening existing apprehensions, such as privacy/targeting concerns [6].

As native advertising is focused on fitting in with ads' surrounding context, we build on these results and further investigate the effects of ad quality and context relevance. Rather than evaluating effects for the advertiser, such as recall and conversion rates, we here focus on effects for users and ad-serving sites. As in [8], we use 'bad' and 'good' ads to refer to ads that are more, or less, annoying to the user. We use a news headline aggregation case as a context in which perceived credibility is important in site quality perceptions [5] and can be affected by ads [11].

METHOD
We use a two-stage Mechanical Turk experiment, limited to US-based participants. We first establish a set of 'good' and 'bad' ads. In the second step we use those benchmarked ads to measure effects on perceived content/site quality.

Step 1: Perceptions of ad quality in isolation
We collected 45 native ads (see Figure 1) from news headline pages. Each ad was rated by 10 participants, resulting in 450 ratings including 3 open text comments each, from 98 participants in total (median=1; 13 of the 98 participants rated 10 or more). Since credibility, especially company trust, plays an important part in attitudes towards advertising [7], we specifically included trust and familiarity ratings, in addition to rating how annoying the ad was perceived (the latter based on [8]).
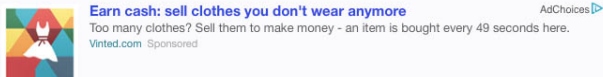

Figure 1 Example of native headline-type ad as rated in step 1

Participants rated the ads on 4 aspects, all 1-5 Likert-type scales. The first 3 items included a 'why?' open text field.

1) In comparison to most ads that you see on websites, how annoying would you say that this ad is?
2) In comparison to most ads that you see on websites, how would you rate the design of this ad?
3) How would you rate the trustworthiness of this ad?
4) How familiar are you with the brand advertised?

Step 2: Ad quality and aggregate content perceptions
The goal of this between-subject experiment is to see to what extent perceptions of news headline aggregations are affected by ad quality. We compare 4 different between-subject conditions, based on the same list of mostly finance headlines accompanied by 2 ads (See Figure 2):

- Good Ads: two native ads rated as not annoying and not untrustworthy in round 1
  1. 'GNR': both good ads but *non-related* to content
  2. 'GCR': both good ads, but one *content-related*: a credit card ad near a credit-related headline. The 2nd ad was non-related, re-used from condition 1.
- **3.** 'B': Bad Ads: same headlines, two native ads rated as both annoying and untrustworthy in round 1.
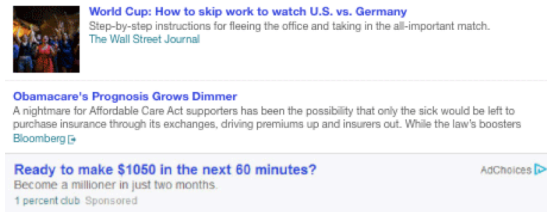- **4.** 'N': No Ads: control, only article headlines.

Figure 2 Example portion (here showing 3 out of 8 items) in ads-content set up step 2. Ads light blue with ad choices logo.

The ads conditions had 6 news stories, 2 ads (8 items). The 'no ads' condition has in total one item less, and one more story from a well-known news agency (7 news stories), avoiding too large of a difference between the number of headlines. Headlines were collected on June 26th, 2014 and kept as-is, including source information. Ads were collected throughout June, and retained their original format with ad demarcation logo and differentiating background color.

Data was collected in three MTurk batches in June 2014. Responses were compensated with $0.30. Items were:

1) How credible do you think this news site is? (1-7: not – very credible, with 'Why' open text field)
2) How would you rate the quality of this collection of news links? (1-7: very low – high, with 'Why' open text field)

Sixty participants rated one of the 4 conditions (n=240 total, 237 valid: 3 participants responded twice, their 2nd response was thrown out). Due to a coding error, in some cases two values could be selected for the credibility rating; the credibility ratings of those participants with two selected values were removed from quantitative analysis (9 out of 237). 214 of the 237 participants of round 2 had not participated in round 1; only 23 (<10%) had rated one or more ads in round 1. Note that they did not necessarily then see those ad(s) in round 2.

RESULTS

Step 1: perceptions of ad quality in isolation
Beyond delivering a set of rated ads, this step provided a number of insights. The most annoying ad was rated with a mean 4.8 out of 5. All questions were correlated (Spearman rho, p<.001, non-normal distributions), and should not be taken as isolated factors. Ads that were seen as annoying, were also seen as less trustworthy (rho=-.525, p<.001), and less attractive (rho=-.468, p<.001). Design attractiveness and trust are positively related as well (rho=.467, p<.001). Familiarity is negatively correlated with annoyance (rho=-.299, p<.001) and positively related to attractive design (rho=.234,p<.001) and trust (rho=.578,p<.001). Ads with an image were rated as more attractive (median 3>2, M-W U=8317.500, p<.001).

Our qualitative findings support findings about the importance of corporate credibility on ad attitudes [7]. The most annoying ads were direct response/sweepstake–type ads, cheaper car insurance, and dating sites. The least annoying were predominantly well-known brands. It is noteworthy that unknown brands and 'make money' type messaging isn't necessarily seen as annoying - as long as claims are credible. One of the least 5 annoying rated ads was about 'making money', from a brand rated as unfamiliar, but offered the believable option to 'sell clothes that you don't want anymore'. Note that perceptions can vary; while on average people may find particular well-known brands trustworthy "…It is a big, well known and trusted company." Others think, "[same brand] is bad". Naming variations of the same brand matter; while two ads for the same consumer appliance brand featuring exactly the same brand name, but with different content scored exactly the same mean (4.3, ranked 5th/45 ads). Two ads from a financial company however differed, the full brand name ad was rated 1st/45 with a mean 4.7/5, while the ad with its abbreviated name scored lower (3.7, midrange) with one participant explicitly stating: "I don't like how it says [short brand name] instead of [full brand name]").

Step 2: ad quality and aggregate content perceptions
Participants in all groups vary in their ratings for the set of headlines, each condition received both the lowest (1) and highest (7) scores on both quality and credibility. As the data is non-normally distributed, we perform non-parametric tests between conditions. There are significant differences for both credibility (Kruskal-Wallis H=9.083, df=3, p=.028) and quality (H=14.997, df=3, p=.002). We follow-up with Mann-Whitney tests, and apply the Bonferroni-Holm adjustment to required significance thresholds to correct for multiple comparisons (Table 1). Note that 1-tailed significance levels are used to reflect expectations on ad quality effects.

Our data shows that having no ads is better than good, content-related ads (credibility U=1153.5, p (1-tailed)=.0025 or quality U=1156, p=.0005), but *not* significantly different from having good, unrelated ads. It's noteworthy we find no significant differences between bad ads and related good ads. No ads is better for perceived site quality than bad ads (quality U=1222, p(1-tailed)=.0015), but has no conclusive results for the credibility of the news site, implying that people distinguish between credibility of news stories and quality of the site overall.

These results imply that beyond ad quality, content-relatedness may play an additional role. This is interesting as it could suggest that a clear distinction between stories and ads is more important than ad quality per se. Indeed, good non-related ads are better for quality perceptions than good related ads (U=1325, p=0.0115) – note that this is unilateral 1-tailed, with one specific type of ad (credit card story next to credit card ad) so further studies are needed to better understand effects, especially for credibility.
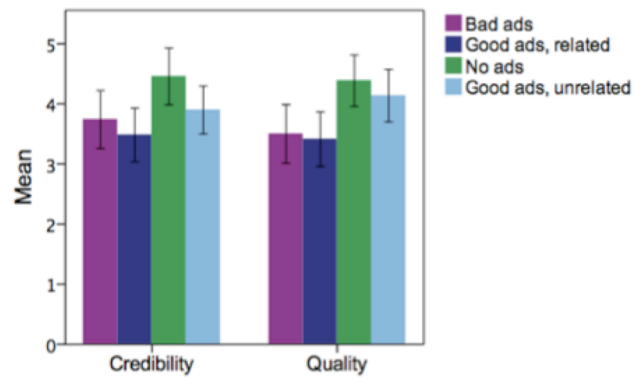

Figure 1 Credibility and quality means per condition. Scores on 1-7 scales. Error bars are 95% Confidence Intervals.

conditions	Credibility		Quality	
	U	p (1tail)	U	p (1tail)
No > GCR	1153.5	0.0025*	1156	0.0005*
No > Bad	1198.5	0.021	1222	0.0015*
GCR < GNR	1460.5	0.083	1325	0.0115*
Bad < GNR	1518.5	0.332	1388.5	0.027
No > GNR	1360.5	0.036	1572.5	0.143
Bad < GCR	1437.5	0.224	1678	0.366

Table 1 Mann-Whitney comparisons between conditions, ordered for Holm-Bonferroni adjustment: target p value/significance n-rank of pair +1. *denotes significance with adjusted 1-tailed .05 threshold.

Participant comments
Participants' comments were used to gain further insight in the found differences. Analysis focused on mentions of advertising/sponsored content and provenance. Comparing the three conditions that include ads (N=177), participants indeed mentioned advertising related terms (e.g. ad(s), sponsor(ed)) most in the *good content-related* ads conditions 36/59 vs 26/59 for bad ads and 23/59 for good non-content related ads. A chi-square test ($\chi2$=6.292, df=2, p=.043; Cramer's V=.189, p=.043) indicated that salience of ads is related to the type of ads.

We see indications that ad quality can affect perceptions "I think it's credible because [credit card brand] is one of the advertisers and [credit card brand] is a very reputable company…"-GCR. Some people expect content-related ads and note when this is not the case: "The ads and the headlines had almost nothing to do with each other."-GNC. However, most comments rather focused on the amount of ads, or the (dis)similarity between ads and content: *"They have some ads on there, but at least they clearly mark them, so that's good. I'm ok with that" -GCR, "Far to[sic] many ads, and they're also a little difficult to distinguish from articles." -GNC. Ads also raised doubts for some about

other content "Too many ads, so I'm wondering if certain organizations pay to have their stories featured." -GNC.

Source'-related terms (e.g. source(s)/reference/author/sites) were mentioned by 83 participants (bad ads 20/59, no ads 19/60, GNC 28/59, GCR 16/59). Comments included: "A good variety of quality sources" -GCR, "This news site seems pretty credible because the articles are all from credible sources like The Wall Street Journal and The Associated Press." -B. This indicates that provenance is important for credibility. We could however *not* find a significant relationship between conditions and the salience of source ($\chi2=5.959$, df=3, sig=.114, V=.159, sig=.114).

The content used here was a mix of hard news and entertainment, with participants not necessarily agreeing on the content itself. Mentions were made of 'real news', 'clickbait', e.g. "It has a variety of topics and there's more 'important' links about actual news than there are links like 'impulse purchases you regret.'"-B, "The articles look like ads I have seen before that tend to come up all the time"-GNC. Interestingly, 10 of the 60 participants in the no ads condition still mentioned ads, "It mixes together seemingly real news articles with obvious paid advertising type links." This supports the notion that potential concerns are pre-existing, and amplified by certain ad types.

CONCLUSION & DISCUSSION

Our results hint that native ads that in isolation have been rated as high quality could still have a negative effect on site credibility and perceived quality if they are too content-relevant. Interestingly, we find that in specific circumstances good, content-related ads could have the same effects as bad ads; while no ads did not do better than high quality, non-related ads. A potential explanation is that content-related ads may cast user doubts over the distinction between content and ads. This could lead to perceptions of lower quality, perhaps compounded by entertainment-style as opposed to 'hard news' headlines.

When we combine our results with those of De Sa et al. [4], who found that content-relevant ads did not provide for a better user experience, one possible conclusion is that focusing on personal relevance may carry less risk for ad-serving platforms than content-relevance. We however also note suggestions in [6], that as soon as concerns have been raised by ad obtrusiveness, existing concerns surrounding ads are amplified. Somewhat similar to their results for other types of advertising strategies, we find that strategies that individually work well, in tandem can have negative effects. A holistic view of combined ad and site quality, as well as considering both long- and short-term effects, are crucial. Users may grow weary of clicking on certain types of 'click bait' content and ads over time. Users will potentially return less, especially when questions on provenance arise. Lower perceptions of site quality can also be a detractor for high value brand advertisers. This is why an informed approach to sponsored content, the interaction of different ad properties and formats and (long-term)

metric choice is extremely important for ad-supported services.

There are clear limitations to this study. We look at one specific type of ad format and content-relatedness. One specific credit-related ad and content combination caused concerns; indeed, noticing the negative effect of this high-quality ad in a pre-version of this experiment, prompted further exploration as presented here. This experiment will have to be repeated in different contexts, with a multitude of ad formats, and content types to provide conclusive insight. We did not look at engagement effects for millions of actual users. However, we provide a signal that it is in the best interest of ad-serving platforms to ensure that ads are not only high quality, but also easily distinguishable from content to avoid doubts on content provenance.

REFERENCES

1. Agarwal, L., Shrivastava, N., Jaiswal, S., Panjwani, S. Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising, *Proc. SOUPS'13.*

2. Buscher, G., Dumais, S., Cutrell, E. 2010. The good, the bad, and the random: an eye-tracking study of ad quality in web search. *Proc. SIGIR'10.*

3. Cho, C-H. Cheon, H. Why do people avoid advertising on the Internet? *Journal of Advertising* 33, 4 (2004).

4. De Sa, M., Navalpakkam, V., Churchill, E. (2013). Mobile advertising: evaluating the effects of animation, user and content relevance. *Proc. CHI'13*, 2487-2496.

5. Gladney, G., Shapiro, I., Castaldo, J. Online editors rate web news. *Newspaper Research Journal* 28, 1, 2007

6. Goldfarb, A, Tucker, C. Online Display Advertising: Targeting and Obtrusiveness, *Marketing Science* 30, (3), 389-404 (2011).

7. Goldsmith, R., Lafferty, B. Newell, S. The Impact of Corporate Credibility and Celebrity Credibility on Consumer Reaction to Advertisements and Brands, *Journal of Advertising* 29, 3 (2000), 43-54

8. Goldstein, P., McAfee, S. (2013) The Cost of Annoying Ads, *Proc. WWW'13*, 459-470.

9. IAB, The Native Advertising Playbook, http://bit.ly/1gzLtDh, accessed Sept 10, 2014.

10. IPG Media Lab, Native Ad Research from IPG & Sharethrough Reveals that In-Feed Beats Banners, http://bit.ly/1qmrhMs, accessed Aug 28, 2014.

11. Lacy, S. Fico, F. The link between newspaper content quality and circulation. *Newspaper Research Journal*, 12,2 (1991) 46-57.

12. Yan, J., Liu,N., Wang,G., Zhang, W. Jiang,Y. and Chen, Z. How much can behavioral targeting help online advertising?, *Proc. WWW'09*, 261-270.