How Dependable are "First Impressions" to Distinguish between Real and Fake News Websites?

Dongchen Huang
Department of Computer Science
Wellesley College
dhuang3@wellesley.edu

Yige Zhu
Department of Computer Science
Wellesley College
yzhu@wellesley.edu

Eni Mustafaraj
Department of Computer Science
Wellesley College
emustafaraj@wellesley.edu

ABSTRACT

In an increasingly information-dense web, how do we ensure that we do not fall for unreliable information? To design better web literacy practices for assessing online information, we need to understand how people perceive the credibility of unfamiliar websites under time constraints. Would they be able to rate real news websites as more credible and fake news websites as less credible? We investigated this research question through an experimental study with 42 participants (mean age = 28.3) who were asked to rate the credibility of various "real news" (n = 14) and "fake news" (n = 14) websites under different time conditions (6s, 12s, 20s), and with a different advertising treatment (with or without ads). Participants did not visit the websites to make their credibility assessments; instead, they interacted with the images of website screen captures, which were modified to remove any mention of website names, to avoid the effect of name recognition. Participants rated the credibility of each website on a scale from 1 to 7 and in follow-up interviews provided justifications for their credibility scores. Through hypothesis testing, we find that participants, despite limited time exposure to each website (between 6 and 20 seconds), are quite good at the task of distinguishing between real and fake news websites, with real news websites being overall rated as more credible than fake news websites. Our results agree with the well-known theory of "first impressions" from psychology, that has established the human ability to infer character traits from faces. That is, participants can quickly infer meaningful visual and content cues from a website, that are helping them make the right credibility evaluation decision.

KEYWORDS

real news websites, fake news websites, website credibility, experimental study, first impression, advertising, halo effect, web literacy

ACM Reference Format:
Dongchen Huang, Yige Zhu, and Eni Mustafaraj. 2019. How Dependable are "First Impressions" to Distinguish between Real and Fake News Websites?. In *30th ACM Conference on Hypertext and Social Media (HT '19), September 17–20, 2019, Hof, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3342220.3343670

HT '19, September 17–20, 2019, Hof, Germany

ACM ISBN 978-1-4503-6885-8/19/09…$15.00
https://doi.org/10.1145/3342220.3343670

1 INTRODUCTION

Over the past decade, the process of finding and consuming news has been continuously shifting away from printed newspapers[1] toward search engines and social networks on the Web. Communication researchers have described this shift as the move away from "direct discovery" (in which people have a direct connection with a particular newspaper) toward "distributed discovery" (where people find news through digital intermediaries such as Google or Facebook) [22]. For example, in the United States, 51% of the audience used social media weekly to find news in 2017, although this proportion declined to 45% in 2018 [13]. Meanwhile, 93% of adults in the U.S get at least some news online (direct discovery and distributed discovery together).[2]

By making use of "distributed discovery", readers are more likely to land on a news publisher's website that they have never heard of, which prevents them from using familiarity with the source itself to judge the quality of the information.

How do people judge what constitutes reliable news? This question becomes more challenging when considering online news, which, with its combination of traditional news and other anonymous sources, makes it difficult for people to develop trust in it [7]. Previous studies in communication research have linked this question to the concept of perceived credibility, which is composed of three parts: medium credibility, site credibility, and message credibility [6]. One can argue that one reason that fake news websites were so successful in 2016 is that they took advantage of the credibility of the medium, in this case, social media, in which ranking is based on social recommendations fueled by the "wisdom of crowds". As the reasoning went, primarily meaningful stories endorsed by friends should be able to raise to the top of the feed. This reasoning was indeed embedded in Facebook's recommendation algorithm, which favored links shared by friends and family compared to those from traditional news publishers.[3] However, hyperpartisan, click-baity, and exaggerated stories that pleased certain audiences were able to benefit from this algorithmic change, spreading widely and eroding the trust in the capabilities of Facebook to deal with misinformation.

As the public becomes more skeptical and less trusting of news discovered on social media[13], the importance of accurately assessing "site credibility" increases. Earlier research on website credibility has established that people often make judgments based on visual characteristics. For example, in a survey of website credibility

[1] http://www.niemanlab.org/2018/05/the-scariest-chart-in-mary-meekers-slide-deck-for-newspapers-has-gotten-even-a-teeny-bit-scarier/
[2] http://www.journalism.org/fact-sheet/digital-news/
[3] Facebook adjusts News Feed to favor friends and family over publishers - https://www.theverge.com/2016/6/29/12055124/facebook-news-feed-algorithm-changes

involving over 2,500 respondents, nearly half of the open-ended comments mentioned the "design look" of a website [8]. Subsequent studies have confirmed that the layout or physical appearance of a website was the most commented when it came to self-reports of credibility considerations [18]. There are also studies suggesting that in the process of credibility evaluation, people are better at recognizing visual cues compared to textual cues [14].

However, what remains looming on the background are constraints in our time and attention. If we click on a link and land on a website, how quickly and reliably can we formulate an intuitive judgment of whether to trust or not the content on this website? Here is where we draw from the field of psychology for answers.

One of the compelling theories developed in psychology is that of "first impressions". Advancers of this theory have empirically determined that 100ms are often sufficient for inferring character traits from an unfamiliar face [24]. Judgements derived based on first impression can often have far-reaching consequences. For example, some experiments have demonstrated that election results can be predicted based on first impressions [21]. Do people demonstrate this "intuitive and unreflective" behavior when faced with unfamiliar websites? That is, do their "first impressions" of a website accurately inform their perception of credibility? And what happens if they have to make quick judgments for many websites in a row, simulating the process of scrolling through a social media feed for news and briefly clicking through to the website to get a sense of whether a news source can be trusted?

This question of how people assess site credibility under time constraints, and sequential context-switching from one website to another, motivates the experimental study we conducted and are presenting in this paper. Furthermore, considering the financial pressures of many news organizations, and the proliferation of many fake news sites for purely financial benefit through advertising technology [2], we introduce another variable in the study, the presence of ads on a website, in order to gauge their effect on users' perceptions of site credibility.

Our study uses mixed methods to analyze quantitative data (perceived credibility scores assigned to websites in different experimental conditions) as well as qualitative data (the justifications that participants provided for their credibility scores). By asking participants to describe their credibility assessment criteria before and after the assessment task, we can compare the prevalence of such criteria. Our results indicate that the credibility signals referenced by users, split roughly equally between signals derived from the textual content and visual aspect of the website.

While the study of perceived credibility of online content has been a focus of research since the early 2000's [5], [11], such studies rely mostly on off-site, large-scale surveys in which participants can use as much time as they want to answer. To our best knowledge, our study is the first to use a controlled experiment setting to test various aspects of the "first impression" theory on website credibility, by varying time exposure and presence of advertising. We find that the majority of our participants are capable to distinguish between real-news websites and fake-news websites based on credibility cues picked up on the fly. However, participants did better with real news websites than with fake news websites. In fact, a few fake news websites were rated as equally credible as real news websites. Given that creating fake news sites is a cheap operation, even when only a few of them succeed, the consequences for our public discourse might be dire. Therefore, we use the results of our study to suggest concrete web literacy practices, which might be useful in educating the public to become more accurate in the task of distinguishing between real and fake news websites.

2 THEORETICAL BACKGROUND

Psychology, the study of perceptions and decision making, provides researchers, who study how users interact with the Web, with a fertile ground from which to draw inspiration. In the following, we describe some theories based on psychological research that informed us.

The theory of "first impression", which suggests that people infer character traits from faces, is a well supported theory in psychology. Psychologists argue that first impression can be reduced to two dimensions: trustworthiness/valence and dominance, which are essential to help people evaluate the threat of strangers. The trustworthiness dimension focuses on perceived intention to help or harm [20], and the dominance dimension concerns the perceived ability to carry out those helpful or harmful intentions. Judgments made based on first impression can have significant impact on any decision-making process and lead to important consequences in people's life such as judicial outcomes [26]. From the perspective of evolutionary psychology [3], humans developed various strategies and cognitive systems to detect and respond to physical threats, so that their chance of survival and reproduction could be increased. As we move into an information-dense society with a heavy exposure to online information, will people extend the application of similar strategies to non-human entities, in our case, websites? Inspired by the idea that people can infer character traits from human faces in a short amount of time, we designed this study to investigate whether people can infer "character traits" from website "faces" and get insight about the credibility level of a website in a short amount of time.

In Oosterhof and Todorov's [15] two-dimensional model, attractiveness was also measured but found to be largely dependent on the trustworthiness dimension. However, replication studies in recent years have shown that the attractiveness perception is also a fundamental dimension underlying social inferences from faces [19]. Psychologists have thus developed a three-dimensional model with the third dimension named "youthful-attractiveness", which captures how attractive a face is. This dimension was found to be positively correlated with the other two dimensions, which means that faces with higher rating of attractiveness, health, babyfacedness tend to have higher rating on trustworthiness (approachability) and dominance. The inclusion of attractiveness as a factor of social inferences from faces can be seen as a social application of what-is-beautiful-is-good effect, also commonly known as the "halo effect."

The research on "first impression" has its roots in the Elaboration Likelihood Model (ELM) [16]. This model provides an explanation for how stimuli are processed and their impact on attitude change. One major component of the ELM is the central route, which refers to the situation in which the individual receiving the information makes a decision (about attitude change) after evaluating the received messages carefully and thoughtfully. Meanwhile, the second

component, the peripheral route, refers to the situation where the individual makes a decision based on the positive or negative cues or simple inferences invoked by the stimulus [16]. For example, making decisions based on the "halo effect" is an application of the peripheral route.

A recent study suggests that users have weaknesses in their ability to tolerate cognitive biases such as the aesthetic bias (another way to refer "halo effect"), in the information credibility judgment process [25]. Furthermore, education background and experience with web literacy practices do not have a significant impact on improving participants' tolerance for cognitive biases [25]. In the same study, the authors also suggest that users are relatively good at content-based verification strategies (evaluating the credibility based on contents).

We regard "site credibility" as different from "information credibility", because it conveys a judgement about the entity that operates the site. Often the information credibility of a particular news story will be affected by the credibility of the news source (or its stand-in, the website) in which it is hosted. Therefore, in our research, we focus on the question of what factors influence people's evaluation process of websites' credibility. Would people be able to perceive a website's credibility reliably even under time constraints? For human faces in the "first impression" study, researchers found reliable judgements in $0.1sec$ [24]. However, websites are not human faces and it is impossible to get the entirety of a website's homepage at once due to the limitation of the screen size. To account for these differences, we found it reasonable to allow for more time for the so-called "first impression", and vary this time to understand its effect. Concretely, our study operates under three time conditions: 6s, 12s and 20s.

Although researchers in other fields (e.g., information science) don't use the concept of the "halo effect", they still have considered semantically related concepts, such as "peripheral cues" [10], which refer to the visual aspect of a web page (e.g., attractiveness of images or professional design). Moreover, they have connected such cues to Herbert Simon's theory of bounded rationality with the implication that, when people are uncertain about the credibility of a web article, they tend to look for cues that can be easily processed, instead of using cues that require more cognitive effort. Compared to content-based evaluation strategies, visual cues are considered easier to process and do not require as much effort. One additional goal of our study then is to discover to what extent users rely on such visual cues (that we refer to as "credibility signals") and what are some concrete and common examples of such visual cues.

3 METHODS

In trying to understand whether people are good at distinguishing between real news and fake news websites, we do not ask them directly to label websites as one or the other category. Instead, we use the concept of perceived credibility to avoid being suggestive. By asking the users to rate a website with a credibility score from 1 to 7, we are able to better capture the inherent uncertainty and scepticism in this task, than the binary choice between real vs. fake news could. Given that we are undertaking this study in the backdrop of the societal discussion on fake news [9], misinformation [1], and hyper-partisanship [12], our focus is on websites that belong to news sources that span the entire credibility spectrum: from the reliable to the fake ones, including sites with a known political bias, as well as sites that are satire and parody. Our dependent variable is the site credibility and we experimentally manipulate time exposure to a website's screen capture image and presence of ads, our two main independent variables. Additionally, we remove the presence of logos, website name, owner of publication, etc. from the image, to create a situation in which a user cannot use the familiarity with branding as a cue in the credibility assessment. In the following, we provide details about the websites used in the study, the participants, and the study procedures.

3.1 Data: Websites of News Sources

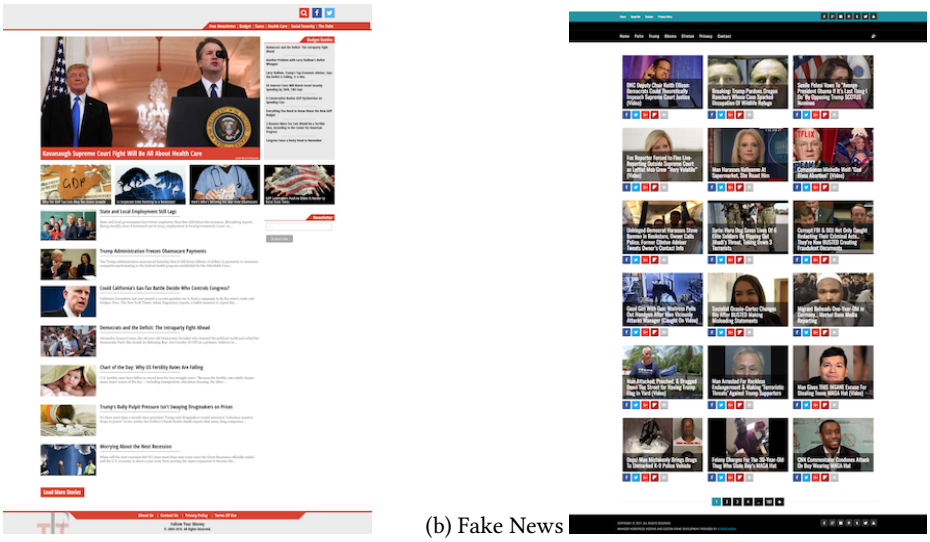
We used 28 news source websites for this study, evenly split between real news and fake news. 14 real news websites were randomly selected from a list of news websites collected from AllSides,[4] a project that uses a combination of techniques to assess the perceived political bias of news sources. These websites belong to real and (relatively) well-known news organizations, that are rated ideologically as center, left, left-leaning, right, or right-leaning by AllSides. An additional 14 websites were randomly selected from a list of low-quality news websites compiled in 2017 by the fact-checking website, Politifact.[5] This list contains websites that are entirely fake news or are imposter sites, websites containing a mix of real and fake news, as well as websites that generate satirical news. In this paper, we refer to the websites selected from AllSides as "real news", and to the websites selected from PolitiFact as "fake news". Both labels are somewhat problematic: "real" news websites that are partisan are often regarded very skeptically, and "fake" news websites that produce satire are given a pass, since their goal is not to inform. Thus, we are using these terms rather broadly and without any judgment, but simply adopting the categories of the list creators, AllSides and PolitiFact. The names of all websites are listed in Table 1.

Real News	Fake News
Associated Press, BBC News, KQED, The Verge, The Daily Targum, The Nation, The Daily Beast, Democracy Now, Deseret News, Richmond Times Dispatch, The Fiscal Times, The Daily Signal, The Washington Free Beacon, Front Page Mag.	Angry Patriot, Awareness Act, bb4sp, Before It's News, Channel 18 News, Chicago Civic Tribune, Daily Snark, Duffel Blog, Empire Herald, India Times, Nation (PK), politicot, The Big Riddle, WestfieldPost

Table 1: The two lists of websites that we used in our study. They were randomly drawn from lists compiled by AllSides and PolitiFact.

As already mentioned, participants did not interact with the real websites. Instead, they were shown a "full page screen capture" that was created using a Chrome plugin.[6] The HTML was modified

[4] https://www.allsides.com/media-bias/media-bias-ratings
[5] https://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/
[6] https://chrome.google.com/webstore/detail/full-page-screen-capture/fdpohaocaechififmbbbbbknoalclacl?hl=en



(a) Real News (b) Fake News

Figure 1: Two screen captures juxtaposed to notice that real news websites and fake news websites often might have a very distinct look-and-feel. All the website layouts were modified to remove their names and logos to avoid recognition.

dynamically prior to the screen capture to remove any names and logos of the website, to avoid the effect of name recognition. For example, Figure 1 shows two examples of screen captures: that of the real news website *The Fiscal Times*, and that of the fake news website *bb4sp*. Screen captures for all websites were created on the same computer (a Mac laptop), with the browser's ad blocker on and off. The browser history was reset before the screen capture process to avoid receiving ads based on past browsing behavior.

3.2 Participants

To carry out this study, we conducted in-person interviews to allow participants to perform a think-aloud explanation of their assessments, but also to comply with the experimentally assigned amount of time to perform the credibility assessment task.

Due to these necessities, our sample is a convenience sample, composed of individuals who we could convince to give us 20 minutes of their time in exchange for a $5 Amazon gift card.

We recruited 42 participants from a variety of environments including a college campus and community gathering spaces. The majority of the participants were female (26), then male (13) and 3 preferred not to answer. They ranged in age from 15 to 66 years old, with a mean age of 28.33 (SD=13.91). The majority of participants had completed a bachelor education or higher (or were in the process or completing), with only 16.7% reporting a high school education. The participants were racially and ethnically diverse (only 21.4% identified as white).

All the interviews were conducted by the first author with the participation and logistic support of the second author, in July 2018.

3.3 Composition of the Study

This study obtained our institution's IRB approval and all subjects signed a consent form before starting the interview, in which among others they agreed to the audio-recording of the interview for the purposes of analyzing its content at a later time. Interviewers did not collect personally identifiable information about the participants. Each interview had a pre-approved and tested structured script composed of the following parts:

(1) **Hypothetical Scenarios:** Three hypothetical scenarios involving the credibility assessment of a website were introduced, one at a time, and participants were asked to describe how they would act under each scenario. The scenarios were as follows:
   (a) Imagine that a friend shared an article with you, containing some interesting data and information, but the article is from an unknown website. What would you do?
   (b) Imagine that you are searching for a knowledge field that you are unfamiliar with. How would you start your research? What websites would you choose?
   (c) Imagine that your professor (or manager) asked you to write a paper (report) about a certain topic, and you decide to cite a certain website as a reference. However, the professor (manager) is skeptical about this particular website, how would you persuade them that the website you chose is a credible source?

(2) **Credibility Definition:** Participants were asked about how they define the credibility of a website, and describe their strategy of determining whether a website is credible by recalling how they came up with the strategies or criteria they mentioned during the previous part of the interview.

(3) **Credibility Rating Task:** Participants were then invited to complete the credibility rating task. Each participant was asked to rate the credibility of eight websites on a 1 to 7 scale, as 1 being not credible and 7 being very credible. The eight websites, randomly selected out of the total 28, consist of 4 real and 4 fake news sites. Participants were randomly assigned in a different time exposure and ads groups, in which they had 6s, 12s, or 20s to scroll through the screen

capture of a website with or without ads. After each rating (captured automatically), participants were asked to explain their credibility assessment score. This can be thought of as a delayed form of a think-aloud explanation for their decision-making. We didn't ask participants to do this during the task, because the time exposure was short and talking and scrolling at the same time would have been difficult. Once the time exposure ended, the browser tab containing a website's screen capture was closed automatically.

(4) **Reflection:** After the rating task, participants were asked to reflect about the completed task and to express any new ideas or criteria that they now deem important for assessing the credibility of websites.

(5) **Demographics Survey:** As the last step, participants were asked to complete a demographic questionnaire that asked for their age, gender identification, education background, race/ethnicity, frequency of checking news online, and three to five websites that they deem credible.

To prepare participants for the limited time exposure for evaluating each website, the credibility rating task contained a test website that was not used for data collection.

3.4 Anonymization of Websites

Although we removed names and logos from all news websites, some "real news" websites may have been familiar to participants, and therefore might have been identifiable from fonts, layout, or contributors. To check the anonymization of the websites, we asked participants manipulation check questions such as "Do you find this website familiar?" or "Do you think you have ever visited it before?". We asked these questions to 34 of our 42 participants. 7 said that they were not familiar with any of the websites used in the study. The rest (27) claimed that they might have visited one or two of the websites before, but could not name the website. Only 6 participants attempted to name 1 out of the 8 websites that they were given, but all failed to have the correct answer. Therefore, we are confident that the anonymization process was successful in concealing the real identities of the websites.

3.5 Limitations

Our study uses a convenience sample. However, we made an effort to go beyond the usual "undergraduate" population used in psychological studies and recruited participants from other settings, too. We did not ask participants about their political affiliation, political knowledge, or their overall news awareness (though we asked to self-report frequency of accessing/looking for news). It is possible that these variables interact with perceptions of credibility for news sources, but our study does not capture that.

This study was conducted using a laptop. Web access statistics indicate that mobile access to web content has surpassed access via desktop-based devices in recent years.[7] Thus, the task of assessing credibility through a mobile device might pose different challenges, which are not captured through our experiment. However, given the lack of widespread ad-blocking opportunities for mobile browsing,[8] users often find themselves in the presence of websites that incorporate ads, a variable for which we tested.

4 RESULTS

Our study collected quantitative and qualitative data, therefore, we used a mixed methods approach for the analysis. Through our credibility assessment tasks, we collected 336 credibility scores for the 6 different experimental conditions. We formulated and tested three main hypotheses through statistical hypothesis testing. Additionally, the 42 interviews with the participants that were audio-recorded during the sessions, were transcribed and manually coded using the approach of "flexible coding" [4].

4.1 Research Hypotheses

These are the three hypotheses that we tested:

- **H1:** Participants will be able to distinguish between "real news" websites and "fake news" websites by assigning higher credibility scores to "real news" websites.
- **H2:** "Real news" websites will be perceived as more credible than "fake news" websites.
- **H3:** The presence of advertisements and time exposure will have an impact on the credibility scores for "real news" websites and "fake news" websites.

The following section describes in detail how we tested each hypothesis and the corresponding results.

4.2 Quantitative Analysis: Hypothesis Testing

The credibility rating task consisted of a between subject, 3 x 2 factorial design with 1) time exposure (6s/12s/20s) and 2) occurrence of advertisements (ads/no ads) as independent variables. Each participant was asked to rate eight screen captures of websites, whose order was randomly generated to eliminate the potential differences in participants' responses that resulted from a fixed order in which the "real news" and "fake news" websites were presented to them. With 28 websites in total, each website was rated by six participants in both ads and non-ads condition, and rated by four participants in every time condition.

4.2.1 H1: Participant Behavior. To test H1, we focused on the 8 credibility scores provided by each participant, averaging by news type (real or fake). These average scores and their error bars are depicted in Figure 2, which indicates that despite the variability of ratings, most participants assigned higher scores to "real news" rather then "fake news" websites. Then, we created a variable called "real-fake difference" by subtracting the mean credibility score of "fake news" websites from that of "real news" websites for each participant. Hypotheses H1 asserts that participants should assign "real news" websites significantly higher scores, which translates into the need for the "real-fake difference" to be significantly greater than 0. As shown in Figure 3, the majority of participants have a positive "real-fake difference". The one-sample t-test for "real-fake difference" indicates that such difference is significantly greater than 0 ($t(41) = 5.09$, $p < .001$), supporting our hypothesis H1. After

[7] https://techcrunch.com/2016/11/01/mobile-internet-use-passes-desktop-for-the-first-time-study-finds/

[8] It is only 5% in the United States: https://digiday.com/media/mobile-ad-blocking-becoming-bigger-threat/

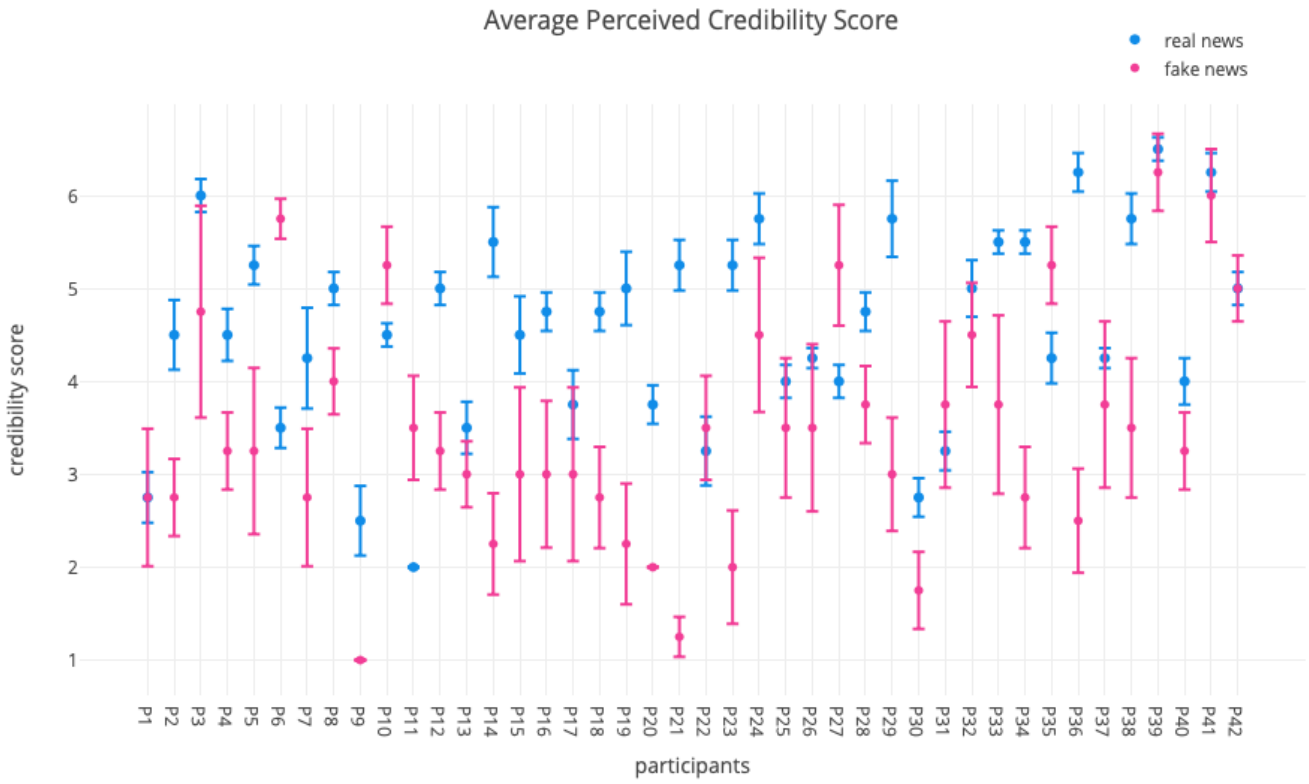


Figure 2: This chart aggregates all average credibility scores that our 42 study participants assigned to "real news" and "fake news" websites. Since each participant rated four of each, the points refer to the average scores for each subgroup and the error line (standard error) is displayed too. Notice the larger error lines for "fake news" websites, indicating the difficulty of assessing their credibility reliably. However, for most participants the real news average credibility scores are higher.

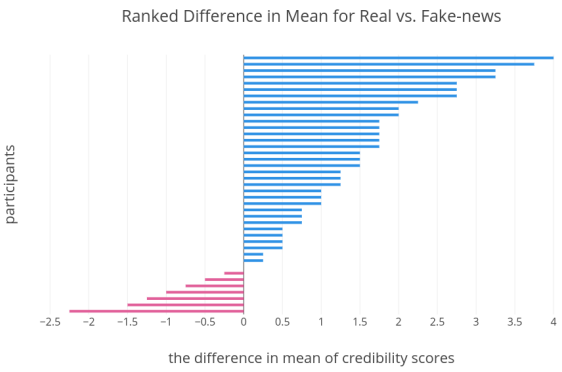


Figure 3: The ranked distribution of the "real-fake difference" variable used in H1 testing. Only 7 participants rated fake news websites as more credible than real news websites, shown by negative values of the difference variable.

testing H1 independently of the two factors in the study (time and advertising), we examined the results subject to these factors. H1 is still supported after controlling for the presence of advertisements or different time setups. For advertisement, the "real-fake difference" is significantly greater than 0 with ads ($t(20) = 3.86$, $p < .001$) and without ads ($t(20) = 3.26$, $p < .01$). For the various time exposures, the "real-fake difference" is significantly greater than 0 under 6s ($t(13) = 2.08$, $p < .05$) 12s ($t(13) = 2.34$, $p < .05$) and 20s ($t(13) = 4.66$, $p < .001$).

4.2.2 H2: Website Perception. In Figure 4, we can observe that scores of "real news" websites ($M = 4.64$, $SD = 1.45$) are clustered at and above the neutral rating (score = 4), and that of the "fake news" websites ($M = 3.53$, $SD = 1.78$) are grouped below the neutral score. More specifically, 54% of the scores of "real news" websites lie between 5 and 7, and only 20% of their ratings are in the non-credible interval, which is from 1 to 3. To the contrary, the majority of the ratings of "fake news" websites (54%) lie in the non-credible interval, and 30% are on the credible side. Also notice that for extreme values, "fake news" websites were rarely rated as highly credible, and "real news" websites were rarely rated as not credible. Figure 5 better captures the nature of H2, by indicating that when averaging over all the scores assigned to a website, there is no overlap between the distribution of averages for the two news groups, except for 1 and 2 outliers in each of them respectively.

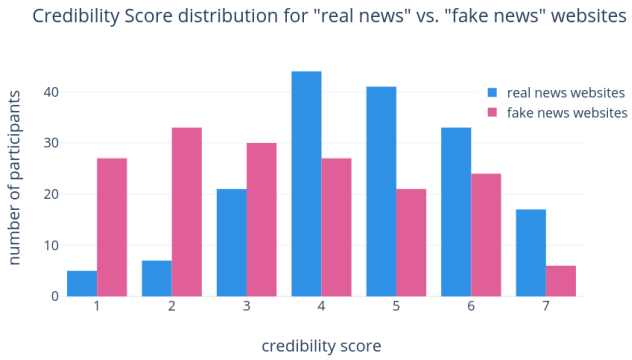


Figure 4: The two credibility score distributions for "real news" and "fake news" are shown side-by-side for comparison purposes. The two distributions are distinct, with the one for "real news" showing (as expected) a higher mean and a more normal-curve shape. Meanwhile, the almost randomly-uniform distribution for "fake news" websites, indicates how difficult it often is to tell them apart from legitimate news sources.

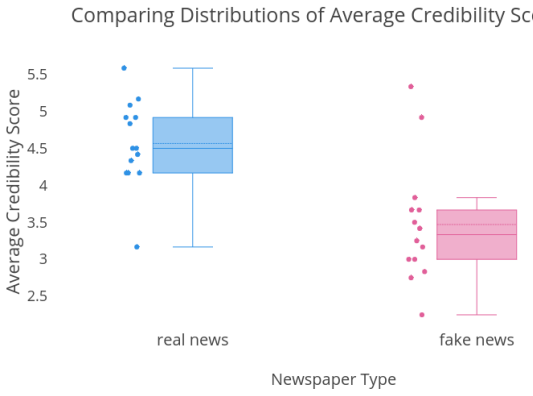


Figure 5: A visual representation for the H2 testing. The boxplots represent the distribution of the average credibility scores for each group of websites. The means of the two groups are 1 point apart. There is one outlier (very low value) in the "real news" group and two outliers (very large values) in the "fake news" group.

To test H2, we conducted multiple two-sample t-tests. Independently of advertisement and time conditions, "real news" websites were rated significantly higher than "fake news" websites ($t(26) = 3.94$, $p < .001$). Being subject to the advertisement condition, "real news" websites still scored significantly higher than "fake news" websites both with ($t(26) = 3.81$, $p < .001$) and without ($t(26) = 3.00$, $p < .01$). The result is consistent under 6s ($t(26) = 1.74$, $p < .05$), 12s ($t(26) = 2.53$, $p < .01$) and 20s conditions ($t(26) = 5.53$, $p < .001$). Therefore, H2 is supported by the study results.

4.2.3 H3: Time and Advertisement. To investigate the impact of time exposure and advertisement presence, we conducted two-sample t-tests, one-way ANOVA and two-way ANOVA tests to examine both the main and interaction effect of these two factors. For advertisement, the average scores for "real news" websites with no ads is significantly higher than that with ads ($t(26) = 1.92$, $p < .05$), and it still holds for "fake news" websites ($t(26) = 2.27$, $p < .05$). This reinforced the finding that advertisements were treated as a negative signal for a website's credibility, encountered in the think-aloud explanations. For time exposure we observed a trending of getting a significant result on scores of "real news" websites ($F(2,39) = 2.65$, $p = .08$). Such a trending could be treated as a signal of a significant positive impact (i.e., being exposed for a longer time, a "real news" website is more likely to get a high credibility score than being exposed for a shorter amount of time). Nevertheless, we can say that only the advertisement portion of the hypothesis H3 was supported by the study results.

After considering the impact of time exposure and ads presence separately, we conducted two-way ANOVA tests to examine the interaction effect of these two vectors. For "real news" websites, the interaction effect of advertisement and time for "real news" websites is not significant ($F2, 78) = .8875$, $p = .42$). This suggests that the impact of advertisement on the score of a "real news" website does not depend on how long the website was exposed to participants. In the group of "fake news" websites, the analysis of variance revealed a significant interaction effect between advertisements and time ($F(2, 78) = 3.81$, $p < .05$). The interaction effect suggests that the impact of advertisements on the credibility score depends on how much "browsing time" the participant received. The negative influence of advertisement becomes more obvious if people were given more time to inspect the screen captures.

4.3 Qualitative Analysis: Credibility Signals

Our Methods section described in detail the various interview questions for the participants. We manually coded the transcripts of the interviews (using the QAD software tool Atlas.ti), by following a new coding protocol, named "flexible coding", outlined in [4]. While this process takes inspiration from the well-known Grounded Theory,[9] it flips the coding process by trying a top-down, instead of a bottom-up process of assigning analytic codes to the text passages of the interviews. The results of our qualitative analysis are very rich and this paper doesn't have space for them. As a result, we will limit our discussion only on the credibility signals that we extracted from the various questions we asked to the participants.

Our study (deliberately) did not use pre-existing definitions of credibility and did not suggest strategies for credibility assessment to the participants. Instead, we sifted through the interview data, in order to extract criteria for the evaluation of credibility from the participants' justifications. In the following, we will refer to these justifications drawn exclusively from the data as "credibility signals".

We distinguish between two groups of credibility signals: pre-task and post-task. Pre-task signals refer to the credibility signals that participants mentioned during the interview (the various hypothetical scenarios), but before the credibility rating task. Post-task

[9] https://en.wikipedia.org/wiki/Grounded_theory

Signals **Textual Content**	**Meaning**	**Quotation**
Sensational Titles	The titles of the article are sensational or click-baity	"some of the titles seem very click baity" "it just seemed very like trying to get your attention by making the most out there headline."
Bias	The articles lean towards one (political) side.	"I think it's is not super credible because it seems like very, it maybe more left leaning." "It did seem like it was leaning more towards one political side than the other,"
Range of Topics	The range of topics shown on the websites	"It had so much information from too many things." "The topics are very scattered."
Entertainment / Gossip	The content is mostly about entertainment or gossip	"It seems like an entertainment magazine. And this type of magazine makes me feel like less credible."
Opinionated	The content expresses opinions rather than giving facts	"I wouldn't call it credible. It's probably opinionated and not so much factual."
Self-identification	A demonstration of contact information	"It had a contact page." "I still saw like privacy policy or other links which might take you to more like explaining the website."
Politics-related	Most of the articles are about politics on the websites	"I'm also just like generally untrustworthy especially when it comes to politics in general."
Inappropriate	The content contains discrimination, pornography etc.	"It looked like a very anti-semitic website."
Recency	The information is up to date	"It looks updated recently."
Visual Aspects		
Images	The quality and quantity of the images on the websites	"I mean there were like joke pictures." "... very carefully selected pictures"
Page Organization	The overall organization of the page, including sections, positions and relative sizes	"Everything is organized in a professional manner." "It's kind of messy and there's a weird ad at the top corner."
Familiarity	If the participant has seen a similar page before	"I mean it looked like BBC." "It kind of reminds me of the New York Times or Huffington."
Ads	The amount of ads on the page	"It lacked credibility in the sense of there's so many ads going on."
Looks and Feels	The gut feeling of participants for the page layout	"I feel like the layout is not convincing." "...plain boring background where you kind of feel like it's just something someone threw together"
Colors	The color theme of the page	"I can't read them obviously but they're just too colorful."
Social Media Icon	The presence of social media icon and share buttons	"The social media links kind of making me skeptical"
Font	The font on the page	"The font was a professional."
Design	The design of the page	"It had like a pretty sleek design"

Table 2: Credibility signals discovered during the manual coding process of the participant interviews. The short description of a signal is shown in the first column, followed by a longer explanation, and quotations of such signals from the interviews. Notice that there are two groups of signals: for the textual content and for the visual aspects of the websites.

signals refer to the criteria mentioned during and immediately after the credibility rating task, i.e. the reason participants gave for their credibility scores. For each mentioned signal, we then aggregated the occurrences during pre-task and post-task sessions.

Once we had identified a large group of signals, we noticed two natural groupings in them: signals related to the textual content of the website, and signals related to its visual aspect. We have shown some of the most frequent signals from each of these two groups in Table 2. Comparing the signals of the post-task group to those of the pre-task group, we found many in the former that were not even mentioned in the latter. This indicates that the evaluation of credibility is contextual and task-specific: although people might have memorized some heuristics that they can usually recall without a context, these may be easily abandoned in favor of present elements in a website that offer new justifications.

This was especially true for the visual aspect group of signals for credibility assessment. Such signals were mentioned only 19 times in the pre-task interviews (that is, they were not easily accessible from memory), but they were mentioned 172 times post-task. Most signals, such as the familiarity with the page layout, colors and font were never mentioned during pre-task questions. And, even though signals such as images and page organization were mentioned a few times pre-task, their significant increase in mentions indicates that people are not aware of the potentially large effect that these signals have on their credibility assessment.

A similar picture emerged for the textual content signals as well. Signals such as "sensational titles", "range of topics", and "opinionated" were barely mentioned pre-task, but people found them useful when they encountered them in the context of the task. Additionally, the topical coverage of a website has a large influence over the credibility judgment: if the website's content covers news, politics, or science, it is viewed as more credible. On the other hand, our participants seem understandably skeptical if they observe the content to be problematic, obviously false, or related to entertainment and gossip. In such occasions, participants mentioned that the content did not look "serious" (for the latter group) or the content looked "in-depth" (for the former group).

The effect of "range of topics" differs among participants. Most participants commented that the website "covers too many topics" or the "topics are very scattered," which made them label the website as not credible. However, some participants said that the variety of different articles made the websites more credible. We noticed that those who attached positive feedback on the range of topics also mentioned that the websites looked more organized. This illustrates the well-known connection between form and content: form helps bring salience to content [23].

5 DISCUSSION

Through our quantitative analysis, we verified our main hypothesis that most participants can distinguish between "real news" and "fake news" websites, because they perceive the former as more credible. During the interviews, we collected participants' definition of credibility and their strategies for evaluating the credibility of online sources. Then, the credibility rating task provided us with a chance to observe participants' real application of their credibility assessment process and contrast it with their self-reported process. In this section, we discuss some of the inconsistencies identified by the comparison of pre-task and post-task credibility signals and the insights we can gain from them.

5.1 Visual signals vs. Content Signals

In the qualitative analysis of interview data, an unexpected pattern was the huge gap between the mention frequency of visual signals during pre-task and post-task. Among all credibility signals mentioned during pre-task, 21% were content related, 21% referred to background information, and only 10% addressed the visual aspect. Meanwhile, during post-task, 47% were content signals and 44% were visual signals. Acknowledging that the rating task is time constrained, it is to be expected that background information signals were hardly mentioned post-task, but is the four-fold increase of visual signals also just a function of limited time? Meanwhile, textual content signals, which take more time to process than visual signals, still weighed the highest among all credibility signals. We hypothesize that the increase in adoption of visual signals might reveal that participants are unaware of the impact of visual signals on their evaluation process of websites' credibility.

Such an unconscious application of visual signals in the credibility evaluation process may also explain the fact that although participants had various definitions of credibility, the rating score indicated a mutual agreement of which type of websites are credible ("real news") and which are not ("fake news"). Compared to content signals and definitions, it is relatively easier to have people agree on visual signals, such as whether a website is appealing, clean, organized, etc. Although this needs more investigation, people seem to have a commonly shared concept of what is a standard website for a news source. Indeed, some participants made their credibility score decision based on the structure of the website and said "It seems like a typical news website, so I think it's credible."

Besides the organization of the website, the visual signal that was referred to the most, was the presence of images on the website. Not only was the quality of images a key factor, but participants also made inferences about the content of a website (e.g. political orientation) from its images. One participant said that "there are some photos of fairly professional people", which was the reason of a high credibility score. However, further investigation is needed to determine the impact of image categories on perceived web page credibility.

5.2 The "Halo Effect"

A direct application of the "halo effect" in our study is the result that the presence of advertisements has a negative impact on the credibility of both "fake" and "real" websites. In other words, removing advertisement from a web page makes it feel "cleaner" and more visually appealing, which leads to a higher credibility score. Meanwhile, the increasing reliance on visual signals that we observed can also be connected with the "halo effect." The key factor of the "halo effect" is that people make decisions based mainly on visual perception, rather than rational analysis. The increment on the mention frequency of visual signals to explain credibility scores provides evidence that people are heavily relying on visual perception, which shares some commonality with the "halo effect."

5.3 Verification Behavior

Responses to the first hypothetical scenario, in which participants were asked to evaluate an unfamiliar website shared by a "friend", indicated that 66% would check the information themselves using various strategies. When it was specified that the information is shared by a close friend, 52% happily switched their position and declared that they trust the article, because they trust the friend. Such a strategy-switching could be regarded as an example of the "cognitive miser" theory, claiming that "human's basic tendency is to default to processing mechanisms of low computational expense" [17]. According to this theory, people prefer practices that require less effort. Indeed, counting on a friend's judgment requires less effort than examining a source in person. Similarly, the shift from content signals toward the use of more visual signals during the credibility rating task might be another example of the "cognitive

miser" theory. However, such a hypothesis requires further investigation, to determine whether reliance on visual signals results from people preferring a less effortful path in the credibility evaluation process, or because people believe that they can derive constant and reliable information about credibility from visual cues.

6 CONCLUSIONS AND FUTURE WORK

This experimental study was designed to understand how humans rate the credibility of unfamiliar online news sources with limited time. Our results suggest that even when short in time, most people can distinguish relatively reliably between "real news" and "fake news" websites, which bears similarity with the "first impression" theory predictions. Furthermore, "real news" websites are in average rated significantly higher than "fake news" ones. Meanwhile, contrary to their self-reported verification strategies, participants made almost equal use of textual content and visual aspect-based credibility signals. Our results seem to suggest that although users are using credibility signals drawn from a website's visual aspect, they might not be aware of such reliance. Thus, an important lesson to draw from our findings is to make this reliance explicit, so that users are aware of their cognitive shortcuts.

6.1 Lessons for Web Literacy

The results of this study suggest several potentially useful lessons for web literacy, which will need to be validated within the framework of web and news literacy programs.

- When visiting an unfamiliar website for the first time, allocate more time to its perusal. Distinguishing between "real news" and "fake news" websites with little time is often challenging for some users.
- Although ad-blockers on browsers allow us to avoid annoying and disrupting ads, when visiting an unfamiliar website it is wise to turn the ad-blocker off in order to notice the quantity and quality of ads on a page. They are often an important signal of "fake news" websites.
- Just because a website employs the familiar structure of a news website, it doesn't mean that it is a "real news" website. Ask yourself why you're trusting a website. If it is only because it looks like a news website, be suspicious and try to find another reason.

6.2 Future Work

The data that we collected through these 42 in-person interviews lend themselves to multiple analyses. We can imagine that in the future we can extend our findings by pursuing other directions.

First, we could scale up this experiment using a crowd-sourcing platform, in order to increase the sample size to better observe the impact of time exposure. Such experiment would not include interviews. Second, in order to understand the negative impact of ads, we could code the appearance of the ads with various features (frequency of ads, positioning, etc.) to try to understand the nature of their effect. Finally, since our manual coding focused mostly on frequently mentioned signals, a more detailed analysis could reveal a more exhaustive list of such signals that can become basis for other experimental studies.

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation through the grant IIS 1751087. The authors are grateful to the Wellesley Cred Lab members for their support.

REFERENCES

[1] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. *Trends in the diffusion of misinformation on social media.* Technical Report. NBER.
[2] Joshua A. Braun and Jessica L. Eklund. 2019. Fake News, Real Money: Ad Tech Platforms, Profit-Driven Hoaxes, and the Business of Journalism. *Digital Journalism* 7, 1 (2019), 1–21.
[3] David Buss. 2015. *Evolutionary psychology: The new science of the mind.* Psychology Press.
[4] Nicole M Deterding and Mary C Waters. 2018. Flexible coding of In-depth interviews: A twenty-first-century approach. *Sociological methods & research* (2018).
[5] Andrew J Flanagin and Miriam J Metzger. 2000. Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly* 77, 3 (2000).
[6] Andrew J Flanagin and Miriam J Metzger. 2007. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society* 9, 2 (2007), 319–342.
[7] Richard Fletcher and Sora Park. 2017. The Impact of Trust in the News Media on Online News Consumption and Participation. *Digital Journalism* 5, 10 (2017).
[8] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences.* ACM, 1–15.
[9] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
[10] Sook Lim and Christine Simon. 2011. Credibility judgment and verification behavior of college students concerning Wikipedia. *First Monday* 16, 4 (2011).
[11] Miriam J Metzger, Andrew J Flanagin, Keren Eyal, Daisy R Lemus, and Robert M Mccann. 2003. Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Annals of the International Communication Association* 27, 1 (2003), 293–335.
[12] Henry A Nasrallah. 2018. The toxic zeitgeist of hyper-partisanship: A psychiatric perspective. *Current Psychiatry* 17, 2 (2018), 17–18.
[13] Newman Nic, Richard Fletcher, Antonis Kalogeropoulos, David A. L. Levy, and Rasmus Kleis Nielsen. 2018. *Reuters Institute Digital News Report 2018.* Reuters Institute for the Study of Journalism.
[14] Thomas Nygren and Mona Guath. 2019. Swedish teenagers' difficulties and abilities to determine digital news credibility. *Nordicom Review* 40, 1 (2019), 23–42.
[15] N. N. Oosterhof and A. Todorov. 2008. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences* 105, 32 (2008), 11087–11092. https://doi.org/10.1073/pnas.0805664105
[16] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion.* Springer, 1–24.
[17] Keith E Stanovich. 2018. Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking & Reasoning* 24, 4 (2018), 423–444.
[18] S Shyam Sundar. [n. d.]. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility* ([n. d.]).
[19] Clare AM Sutherland, Julian A Oldmeadow, Isabel M Santos, John Towler, D Michael Burt, and Andrew W Young. [n. d.]. Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* 127, 1 ([n. d.]).
[20] Alexander Todorov. 2008. Evaluating faces on trustworthiness: an extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences* 1124 (2008), 208–24.
[21] Alexander Todorov, Anesu N Mandisodza, Amir Goren, and Crystal C Hall. 2005. Inferences of competence from faces predict election outcomes. *Science* 308, 5728 (2005), 1623–1626.
[22] Benjamin Toff and Rasmus Kleis Nielsen. 2018. "I just Google it": Folk theories of distributed discovery. *Journal of communication* 68, 3 (2018), 636–657.
[23] Bill VanPatten. 1990. Attending to form and content in the input: An experiment in consciousness. *Studies in second language acquisition* 12, 3 (1990), 287–301.
[24] Janine Willis and Alexander Todorov. 2006. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science* 17, 7 (2006), 592–598.
[25] Yusuke Yamamoto, Takehiro Yamamoto, Hiroaki Ohshima, and Hiroshi Kawakami. 2018. Web Access Literacy Scale to Evaluate How Critically Users Can Browse and Search for Web Information. *Proc of WebSci '18* (2018).
[26] Leslie A. Zebrowitz and Susan M. Mcdonald. 1991. The impact of litigants baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior* 15, 6 (1991), 603–623. https://doi.org/10.1007/bf01065855