The Goldilocks zone: young adults’ credibility perceptions of online news articles based on visual appearance

Jacob O. Wobbrock, Lara Hattatoglu, Anya K. Hsu, Marijn A. Burger & Michael J. Magee

To cite this article: Jacob O. Wobbrock, Lara Hattatoglu, Anya K. Hsu, Marijn A. Burger & Michael J. Magee (2021): The Goldilocks zone: young adults’ credibility perceptions of online news articles based on visual appearance, New Review of Hypermedia and Multimedia, DOI: 10.1080/13614568.2021.1889690

To link to this article: https://doi.org/10.1080/13614568.2021.1889690

Published online: 28 Feb 2021.
The Goldilocks zone: young adults’ credibility perceptions of online news articles based on visual appearance

Jacob O. Wobbrock, Lara Hattatoglub, Anya K. Hsu, Marijn A. Burger and Michael J. Mageec

The Information School | DUB Group, University of Washington, Seattle, USA; bDepartment of Human Centered Design & Engineering, University of Washington, Seattle, USA; cThe Information School, University of Washington, Seattle, USA; dDivision of Computing & Software Systems, University of Washington, Bothell, USA

ABSTRACT
Credibility judgments of online news are affected greatly by perceived expertise and trustworthiness, but users encounter an article’s visual appearance before its content, and yet visual appearance has not been studied in isolation. We conduct two studies of news article visual appearance. The first was with 31 undergraduates who rated the credibility of synthetic newslike articles containing only “lorem ipsum” text, indistinct videos and images, non-functional hyperlinks, and various fonts. The second study was with 30 different university students who rated the credibility of news articles from popular web outlets, half credible and half not. The articles were presented at 5600 words per minute, or 20 times faster than typical reading speeds, enabling only judgments of appearance, not substance. Findings show that credibility is affected by article length, image count and density, and font face and size. These factors interact to yield differential effects on perceived credibility. Articles that struck a balance among factors were most credible, giving rise to the notion of a “Goldilocks zone”, where credibility is highest. Interviews from both studies also revealed that perceived credibility was highest for articles that struck a balance among factors. This work has implications for visual information design, especially for online news.

ARTICLE HISTORY
Received 27 May 2020
Accepted 9 February 2021

KEYWORDS
online news; fake news; credibility; trust; content; visual appearance; visual design; information design

Introduction

The prevalence and consumption of online news, including so-called “fake news” (Lazer et al., 2018), has been an undeniable influence on contemporary life and society. Web-based news articles have immense flexibility in their appearance and design, and with today’s authoring tools, far more people can create and publish “news” sites than can produce or distribute printed newspapers. The influence of web-based news was seen clearly in the spread
of fake news promulgated by social media during the 2016 U.S. presidential election (Allcott & Gentzkow, 2017; Grinberg et al., 2019). American citizens, and others, widely shared misleading or false articles because they looked convincing and affirmed their viewpoints or biases.

University students are among the heaviest users of social media and online news in the United States, and social media is a key conduit to news for many students. The hyperlinked nature of social media, and the ease with which news can be shared (often before it has been read), means that online news can proliferate rapidly within this demographic. According to Pew Research (2019), college-educated Internet users are approximately 50% more likely to use social media than their peers, and users between the ages of 18 and 29 are almost twice as likely as any other demographic to use social media. In Europe, younger adults also rarely read news in print, preferring online sources (Matsa et al., 2018). Given the significant role of university students in promulgating news via social media, there is a need to better understand this demographic and their perceptions of news on the Web.

Many factors contribute to the proliferation of web-based news articles, including their content, provenance, timing, appearance, and easy hyperlinked shareability. Although much has been made of web page content (Flannagin & Metzger, 2000; Quandt, 2008; Wathen & Burkell, 2002), much less is known about how the appearance of online news sources might contribute to their perceived credibility. Moreover, the appearance of news web pages is, as with all web pages, instrumental in forming users’ first impressions (Lindgaard et al., 2006; Tractinsky et al., 2006). And as recent research shows, users often only rely on their first impressions, never investigating fake news when it appears in their social media feeds (Geeng et al., 2020), even sharing it without reading it. This behavior is in keeping with prior research showing that users form credibility judgments first based on prominent surface features (Fogg, 2003; Wathen & Burkell, 2002). Given these concerns, we sought to understand which presentational factors affect the perceived credibility of newslke web pages, and how.

To achieve this understanding, we conducted two complementary studies, each of which solicited participants’ credibility perceptions by isolating article presentation from article content in two different ways, thereby triangulating our results. In the first study, 31 undergraduates interacted with a series of newslke web pages generated by a system we built called Pyrite. These generated “articles” were intentionally devoid of any meaningful content (see Figure 1). Specifically, all text generated was from the first-century B.C. “lorem ipsum” speech by Cicero (Wikipedia, 2017); videos and images were mere swirls of blurry colours; and hyperlinks were all “dead” (i.e. non-functional). We conducted a study in which we controlled or randomised presentational factors, including font sizes and serifs, the number of images and videos, video placement, the number of words in the article, and the density
of hyperlinks. We had the 31 undergraduate participants rate on a 1–7 scale how believable each of 24 newslike web pages would seem if they had had actual content.

In the second study, 30 university students, none of whom participated in the first study, rated the credibility of actual news articles taken from popular web outlets. Half of the articles were real news and half were fake news from questionable or satirical sources, unbeknownst to the participants. Unlike in the first study, these articles had content; however, the articles were presented extremely quickly using a rapid exposure protocol similar to those commonly used in psychology to prevent deeper processing (Potter, 1984). Specifically, all articles were presented at a reading rate equivalent to 5600 words per minute—about 20 times faster than typical reading speeds (Carver, 1992; Legge et al., 1985; Seidenberg, 2017; Witty, 1969). Participants could neither read the articles nor examine their words, pictures, or videos; rather, the experience was one of the articles flashing by, leaving only a visual first impression. Participants were shown 100 articles randomly chosen from a collection of 200 articles, rating each article’s believability on a 1–7 scale, as in the first study.
Our central finding from both studies was that certain presentational factors, even when isolated from content, do, in fact, affect the perceived credibility of online news. These factors include article length, image count and density, and font face and size, many of which interact to yield differential effects on perceived credibility. In general, articles that struck a balance among factors were seen as most credible, giving rise to the notion of a “Goldilocks zone”\(^1\) where credibility was highest. Post-session interviews in both studies also revealed that perceived credibility was highest for articles that struck a balance among factors, corroborating our quantitative results and lending support to the existence of such a zone.

The main contributions of this research are our empirical findings from two complementary studies of newslike web pages, each of which isolated presentational factors in different ways, and our conceptualization of the Goldilocks zone. Our first study, previously reported (Wobbrock et al., 2019), had participants rate newslike web pages devoid of meaningful content, giving participants as much time as they wanted on each page. The second study, conducted to triangulate the findings from the first study, had participants rate actual articles containing content, but the articles were presented so rapidly that the content could not be apprehended. Our findings can inform the design of online news, and can also inform citizens as to the effects of presentational factors on their perceptions of news sources. To the best of our knowledge, this work is the first to isolate the appearance of online news from news content in the study of online news credibility perceptions.

**Related work**

In this section, we review work on source and message credibility, web credibility, and fake news and social media. For our purposes, and in keeping with prior work on website credibility (Fogg et al., 2001), we adopt a definition of “credibility” as “believable or trustworthy”. Interestingly, the root word for credibility is from the Latin *credere*, meaning “to believe”, highlighting the link between the two terms. We therefore do not make a distinction between “credibility” and “believability”, allowing, as Fogg et al. (2001) have done before us, for “credible information” to be regarded synonymously with “believable information”.

Prior research has provided in-depth literature reviews of credibility generally and media credibility specifically. For considerations of space, we do not replicate such reviews here, but direct interested readers to prior work (Choi & Stvilia, 2015; Gaziano & McGrath, 1986; Hilligoss & Rieh, 2008; Metzger & Flanagin, 2007; Pornpitakpan, 2004; Rieh & Danielson, 2007).

\(^1\)Goldilocks was the protagonist in the famous nineteenth-century children’s tale, *Goldilocks and the Three Bears*. Goldilocks looks for porridge that is neither too hot, nor too cold, but “just right.”
Source and message credibility

Source credibility has been studied in various contexts. Hass (1981) defined a “credible source” as a source that conveys accurate information, and does so without bias. When evaluated by television news audiences, source credibility has been shown to be based on apparent expertise and trustworthiness (Ibelema & Powell, 2001). Pornpitakpan (2004) provides a thorough review of research on source credibility.

When the source of a message is unknown, evaluations of credibility tend to shift to the message itself (Rosenthal, 1971). Slater and Rouner (1996) showed that message credibility interacts with source credibility to produce overall credibility perceptions. Metzger et al. (2003) argued that message structure, message content, language intensity, and message clarity are the key factors constituting message credibility. Judgments of message credibility are also shaped by the types of information conveyed (Flanagin & Metzger, 2000). Olaisen (1990) showed that factors related to the source or content of a message are distinct from factors related to a message’s medium and its design features, the latter related to what is being investigated here.

Web credibility

Given the ever-increasing amount of information online, it is no surprise that web credibility has been an active topic of research for some time. Writers and researchers have long sought to understand how website credibility judgments are formed. Easy access to web hosting services means that online information is not governed by the same professional gatekeepers as print media, and therefore has a higher risk of being inaccurate (Liu, 2004; Metzger & Flanagin, 2013). Research has shown that Internet users generally lack both the motivation and skills to verify the information they find online (Amsbary & Powell, 2003; Geeng et al., 2020; Meola, 2004). Flanagin and Metzger (2000) showed that even when users do possess the skills to verify information on the Internet, in practice, they rarely bother doing so.

Studies of website credibility have shown that credibility judgments are rapid and complex, incorporating multiple dimensions simultaneously (Fogg & Tseng, 1999). For example, Freeman and Spyridakis (2004) showed that readers evaluate credibility based on objective judgements about the information’s accuracy as well as subjective judgements about the information’s “trustworthiness, expertise, and attractiveness”. Flanagin and Metzger (2007) and Furman (2009) argued that credibility perceptions are based more on visual attributes of a web page, like design features and apparent site complexity, rather than knowledge of the source of the information. Fogg’s (2003) Proximity-Interpretation Theory would agree, holding that web credibility perceptions first depend on prominent visual features before any judgments,
or interpretations, are made. Tuch et al. (2012) also found that visual complexity plays a role in forming aesthetic judgments of websites, and Tractinsky et al. (2006) contended that first impressions of website attractiveness rapidly affect perceptions of trustworthiness. (Beldad et al. (2010) reviewed factors related to website trustworthiness generally.) Although we are unaware of any studies that focus solely on presentational factors as we do here, some prior work has included visual elements among other factors when investigating web credibility. Fogg et al. (2003) conducted a study in which 46.1% of their respondents mentioned looking at the high-level design of a website when forming credibility judgments. As they observed, “No matter how good a site’s content, the visual aspects of a web site will have a significant impact on how people assess credibility”. Robins and Holmes (2008) expanded on this finding, establishing that content with a “higher aesthetic treatment” was perceived as more credible. A literature review by Wathen and Burkell (2002) established that to positively impact perceived credibility, a website must “emphasize a good interface and project a professional image, making use of established design principles”. Stonewall and Dorneich (2016) studied web page appearance, gathering users’ ratings of visual attributes (e.g. colours, shapes, images) as they related to professionalism and gender. Kim and Moon (1998) studied visual attributes in online banking, finding that professional looking graphics and colours increased perceived trustworthiness. Spillane et al. (2017) conducted a crowdsourced credibility study of distorted websites where certain types of content were present or absent (e.g. banner ads, share buttons, comment fields). Our current studies affirm prior high-level findings of the importance of visual appearance on credibility, but our studies go further by isolating presentation from content in newslike articles, and by identifying which presentational factors affect perceived credibility, and how.

**Fake news and social media**

At the turn of the millennium, print newspapers were generally viewed as more credible than online news (Kiousis, 2001). Since that time, many traditional print media have moved online, and by August 2017, 43% of Americans reported getting their news primarily from online news sources (Gottfried & Shearer, 2017). As Burbules (1998) observed over 20 years ago, the traditional reliance on established, reputable news sources for information has been diluted by the Web. The variety of newslike information available online has introduced a kind of “leveling effect” that gives all hyperlinked information, reliable or not, an equal level of visibility.

This trend culminated in the 2016 U.S. presidential election (Allcott & Gentzkow, 2017; Grinberg et al., 2019), during which the term “fake news” was coined due to the proliferation of misleading and inaccurate news articles
via social media (Hunt, 2016; Lazer et al., 2018). These fake news sources crafted volatile and biased stories about presidential candidates and other political figures, and such stories were then shared on social networks even more widely than the most popular mainstream news stories (Silverman, 2016). Surveys conducted at the time indicated that most people who read fake news articles believed them, never verifying the “facts” reported (Silverman & Singer-Vine, 2016). A study by Allcott and Gentzkow (2017) established that the majority of American adults viewed and remembered at least one fake news story during the presidential election, indicating the impact of fake news on the American electorate.

In recent years, the effects of fake news have only been amplified by increased and widespread use of social media. In one study, Geeng et al. (2020) caused fake news to appear in participants’ Facebook and Twitter feeds unbeknownst to them, and found that participants did not investigate fake news thoroughly, or even at all, especially when it came from a trusted connection. Kim and Dennis (2019) found that highlighting the source of a fake news article on social media made users more skeptical of its content, but that users nonetheless were quick to share fake news that aligned with their beliefs. Morris et al. (2012) specifically looked at the credibility of tweets, finding that users generally made judgments based on surface features like message topic, author name, and author image, rather than based on message content. Yang et al. (2013) made similar findings when comparing credibility perceptions of Twitter posts in the U.S. and Weibo posts in China.

Given these and other concerning findings, attempts have been made to help users better apprehend credibility on social media. For example, Schwarz and Morris (2011) augmented web pages and search results with visualisations of additional information to help users assess credibility. Similarly, Yaqub et al. (2020) augmented social media posts with credibility indicators, finding that they helped reduce participants’ propensity to spread fake news. Zhang et al. (2018) developed a set of credibility indicators based on an article’s content and context, the former being internal to an article and the latter being external or based on metadata. And many researchers have attempted to detect fake news automatically through data mining, but doing so remains a difficult problem (Shu et al., 2017). Therefore, some researchers have attempted to use human crowdworkers to perform micro-tasks designed to assess news credibility (Bhuiyan et al., 2020).

Given the impact of fake news, the proliferation of online news via hyperlinked social media, the heavy social media use by young adults, and the importance of visual appearance in affecting users’ credibility judgments, we sought to isolate just how much presentational factors affect credibility judgments of online news. We now turn to the first of our two studies investigating such factors.
Study 1: method

The purpose of our first study was to isolate which presentational factors affected undergraduates’ perceived credibility of newslike web pages. We also investigated how presentational factors affected time-on-page, and conducted interviews after study sessions to understand participants’ subjective perceptions.

Participants

We recruited 31 college students for our first study. Thirteen participants were female (42%) and 18 were male (58%);² the mean age was 20.8 years (SD = 1.4). Three participants were recruited by speaking to a class at our local university and inviting students to participate in our study; the other 28 participants were approached in the university library. Each participant was compensated with $10 USD for about 25 minutes of their time.

As stated above, in our first study, we focused on undergraduate students (typically 18–22 years old) because of their high engagement with social media and online news (Pew Research, 2019). Our participants were therefore limited to undergraduates currently or recently enrolled in college. Therefore, our results might not generalise to people of other ages, or to non-college students of a similar age.

Of our 31 participants, 12 were arts or science majors, six were from information science, six were from engineering, and three were from business. The remaining four were from professional disciplines: education, pharmacy, or public health.

Apparatus

To conduct our study, we built a custom online testbed called Pyrite. Pyrite ran in a web browser and presented newslike “articles” devoid of meaningful content to participants (see, e.g. Figure 1, left), recording their perceived credibility judgments and times-on-page. Pyrite displayed articles in the Google Chrome web browser running in full-screen mode with no other windows open or applications running. Five of 31 participants used Macintosh desktops, desktop mice, and had 27” displays; the other 26 participants used Macintosh laptops, laptop trackpads, and had 13” displays. As we did not conduct any sessions on smartphones or tablets, our findings might not generalise to screen sizes smaller than those of typical desktops or laptops.

Pyrite generated newslike articles based on real-world web page designs employed by actual online news sites. To arrive at the designs, we took the

²Options for gender included “male”, “female”, “non-binary”, and “prefer not to respond”. All participants chose either “male” or “female”.
top 20 U.S. news websites from the Alexa rankings\(^3\) of most popular websites. From there, we visited five random articles from each site’s “news” or “world” sections. From this total sample of 100 news articles, we measured font faces and sizes, word counts, link densities, and image and video sizes, counts, and placements. We then built these values into Pyrite such that it used them as parameters when generating synthetic newslike articles.

As noted above, articles generated by Pyrite were intentionally devoid of meaningful content so that participants would have no content-based influences on their credibility judgments. If we had allowed discernable content into the articles, no matter what the topic chosen, that content could have introduced confounds that would have affected our ability to isolate purely presentational factors. Even innocuous-seeming topics like “cats” would appeal to “cat people” more than “dog people”, let alone actual newsworthy topics like politics or the environment. We recognise the tradeoff in making this choice, namely that we miss the possibility of detecting how presentational factors might interact with web page content. Our second study addresses this issue by presenting actual news articles, albeit very rapidly, rather than synthetic “articles”, as in this study.

Pyrite displayed articles with deliberately manipulated visual features. For text, it drew from Cicero’s well-known first century B.C. “lorem ipsum” text (Wikipedia, 2017), using either 348, 644, or 1070 words. Links were randomly applied according to a link density factor reflecting low, medium, and high density. Similarly, font sizes were chosen for title and body fonts that represented small, medium, and large sizes. Fonts were either serif or sans serif in style. Specific values for these factors are reported below.

For images and videos, Pyrite used heavily blurred media that became indistinct so as not to distract participants with their content (Figure 2). Videos were positioned either at the top or in the middle of articles, but not at the bottom, as we did not observe that placement in our sample of 100 real news articles. Videos were under 60 seconds in duration, but all participants quickly realised that videos were meaningless, and ceased playing them.

**Procedure**

Our study unfolded in three stages. First, we collected basic demographic data about our participants using an online questionnaire. Second, we used Pyrite to show participants a series of newslike “articles”, obtaining their perceived credibility ratings on a 7-point Likert scale, described in detail below. Third, we conducted post-session interviews in which we asked our participants questions about their experiences during the study session.

Participants were shown a randomised series of 24 web page “articles”, some examples of which are shown above in Figure 2. All 31 participants provided

responses for all 24 articles. For each article, participants responded on an agreement-based Likert scale ranging from 1 = “Strongly Disagree” to 7 = “Strongly Agree”. The Likert scale appeared at the bottom of each Pyrite-generated web page. The prompt for the Likert scale was:

*If it had real content, I would believe an article that looks like this.*

The wording of our prompt was influenced by prior work on web credibility (Fogg & Tseng, 1999), which discussed key terminology for investigating credibility, with the top three terms being “credible”, “believable”, and “reputable”. We selected the word “believe” because of concerns that “credible” or “reputable” might lead participants more to consider the authorship or provenance of the article, rather than its presentation. Also, the word “believe” serves as an active verb for the participant.

In addition, our prompt was designed to avoid extracting mere professionalism judgments from our participants (Stonewall & Dorneich, 2016). Our Pyrite-generated web pages, despite their controlled variations, all had a similar “look and feel”. As a result, their level of professionalism was quite similar and unlikely to cause differences in perceived credibility. A more professionalism-oriented prompt might have focused on the niceness of the presentation, rather than on its believability. Our follow-up interviews gave no indication that professionalism was the underlying construct being judged by participants.

Prior to rating web page articles for their perceived credibility, participants did a training exercise that walked them through a practice article. Explanatory prompts were shown as participants scrolled through the practice article (Figure 3), and the study facilitator verbally checked for each participant’s
understanding. Specifically, the first prompt (Figure 3, top) appeared when the practice article loaded; the second prompt (Figure 3, middle) appeared when the practice article had been scrolled down about one-third of the way; and the third prompt (Figure 3, bottom) appeared once the Likert scale at the bottom of the practice article became visible.

After rating all 24 articles presented to them, participants were asked to respond freely in a semi-structured interview to the following questions:

1. What were your first impressions of the pages you saw?
2. How did you evaluate each page for its believability?
3. What elements or characteristics of each page did you find yourself looking at?

Of course, we recognise that participants’ answers to these questions cannot be taken as ground truth for participants’ actual behaviour. It is quite possible, and even likely, that participants did not know what they were looking at or basing their judgments on. Our intention, however, was to discover what participants thought mattered to them in making their credibility ratings, and what they thought they were looking at. Ultimately, ground truth for such questions lies beyond the scope of the current study; future work could pursue such answers with a different study design (e.g. using eye-tracking).

Figure 3. While scrolling through a practice article, participants were shown these three prompts in succession. The prompts were shown in isolation; they are stacked here for considerations of space.
Experiment design & analysis

Recall that the levels of our factors were determined by our survey of 100 articles from 20 of the most popular U.S. news websites. Our study utilised a partial within-subjects factorial design with the following factors and levels administered to all participants in a fully-crossed design:

- **Video**: absent, present
- **Images**: 0, 3, 6
- **Link Density**: 0.000, 0.002, 0.007, 0.017 links per word
- **Trial**: 1–24

For each combination of the levels of the primary factors above, one level was selected randomly for each of the following secondary factors:

- **Words**: 348, 644, 1070
- **Font Size (body/title)**: 13/30, 16/38, 19/46 pt
- **Font Face**: serif, sans serif
- **Video Placement**: top or middle of article

In addition to the above factors and levels, we examined participants’ **Age** and **Gender**. Values for **Age** ranged from 18 to 23 and four options were offered for **Gender**. (See footnote 2.)

We chose the above study design to allow us to investigate the presentational factors that we hypothesised might affect web page credibility based on prior work (Flanagin & Metzger, 2007; Fogg et al., 2003; Furman, 2009; Kim & Moon, 1998; Tuch et al., 2012), while also avoiding making our study impractical to run and analyse. (If all primary and secondary factors had been fully crossed, study sessions would have been impractically long.)

In all, 31 participants each completed 24 trials for 744 total trials. A single trial involved viewing a synthetic web page “article” generated by Pyrite and indicating a perceived credibility rating on a 1–7 Likert scale. The dependent variables for each trial were a **Credibility** ordinal response and **Page Time**, measured in milliseconds. (Participants could explore a page for as long as they liked before marking their credibility rating at the bottom of an article.)

Our data analysis unfolded in stages. With many covariates, factors, and levels, we could not justifiably throw all potential effects into a single statistical model of high order. Instead, as is common practice, we followed a typical variable selection process (Heinze et al., 2018), whereby we first used exploratory data analysis, including descriptive statistics, graphical plots, and outlier analyses, to determine which factors seemed like they might exert a significant effect on either **Credibility** or **Page Time**. We also tested each factor in isolation and in all two-way interactions. Factors involved in statistically significant
(p < .05) or marginal (.05 ≤ p < .10) effects during this stage were preserved in the final statistical model. Factors that did not emerge as potentially significant were not explored further.

For the Credibility measure, the final statistical model contained four fixed effects: Video, Images, Words, and Font Size. This meant that Link Density, Font Face, and Video Placement did not exert a detectable effect on Credibility, and were therefore dropped. Video was encoded as a dichotomous variable while Images, Words, and Font Size were encoded as ordinal variables corresponding to their respective low, medium, and high values. The statistical model also had Trial and Subject included as random effects (Frederick, 1999; Littell et al., 1998).

For the Page Time measure, the final statistical model had three factors: Video, Images, and Words. This meant that Link Density, Font Size, Font Face, and Video Placement did not exert a detectable effect on Page Time, and were therefore dropped. Factor encodings were the same as for Credibility. Again, Trial and Subject were included as random effects.

We ran these statistical models according to established procedures. Specifically, we used the nonparametric Aligned Rank Transform (Higgins & Tashtoush, 1994; Salter & Fawcett, 1985; Wobbrock et al., 2011) to analyse Credibility as an ordinal response. We utilised a parametric linear mixed model analysis of variance (Frederick, 1999; Littell et al., 1998) for the logarithm of Page Time, which was lognormally distributed, as is typical of temporal measures (Lawrence, 1988; Limpert et al., 2001). Statistical tests were conducted in R using the ARTool, lme4, car, phia, and emmeans packages.

**Study 1: results**

In this section, we present the results of our first study of online news credibility perceptions based on presentational factors. We first discuss Credibility and then Page Time. Then we share our interview results.

**Perceived credibility**

Recall that participants each rated 24 synthetic newslike articles containing “lorem ipsum” text on 1–7 Likert scales ranging from 1 = “Strongly Disagree” to 7 = “Strongly Agree” in response to the prompt, “If it had real content, I would believe an article that looks like this”.

An omnibus test shows that there were significant main effects of Video (F_1, 655.4 = 14.87, p < .001), Images (F_2, 654.8 = 13.33, p < .0001), and Font Size (F_2, 667.4 = 6.96, p < .01) on Credibility. By contrast, Words did not exert a detectable main effect (F_2, 667.3 = 2.06, n.s.). However, there was a significant Images × Words interaction (F_4, 665.7 = 2.43, p < .05), and a significant Images ×
Words × Font Size interaction ($F_{8, 644.9} = 2.19, p < .05$). In the following paragraphs, we unpack each of these results.

Figure 4 shows Credibility ratings for when video was absent or present. The mere presence of a video increased the perceived credibility of synthetic articles.

Figure 5 shows Credibility ratings for each level of Images. Interestingly, it seems that three images might be more credible than either zero or six images. *Post hoc* pairwise comparisons using the correction from Tukey (1949) indicate that zero images were significantly less credible than three or six images ($p < .01$), but that three and six images were not detectably different.

Recall, however, that Credibility was affected by a significant Images × Words interaction, as shown in Figure 6.

We can use interaction contrasts (Marascuilo & Levin, 1970), corrected for multiple comparisons with Holm’s sequential Bonferroni procedure (Holm, 1979), to examine the credibility differences between image levels at 348, 644, and 1070 words. Results indicate that the significant difference in credibility between zero and three images at 348 words disappears at 644 words ($p < .05$). This difference does not quite re-emerge at 1070 words ($p = .15$). Therefore, the number of images used significantly affects perceived credibility when the word count is low, but no longer seems to matter as much when the word count increases. Articles with few words that have either no images or many images seem less credible than articles with a medium number of images.
Figure 5. Average credibility ratings by image count. Higher is “more credible”. Error bars and values in parentheses represent +1 SD.

Figure 6. For 0, 3, and 6 images, credibility ratings changed over 348, 644, or 1070 words. Higher is “more credible”. Error bars are omitted for readability.
Figure 7 shows Credibility ratings by level of Font Size. It seems that while small and medium font sizes had similar credibility ratings, large fonts reduced credibility. This result is confirmed by significant post hoc pairwise comparisons using the correction from Tukey (1949), which show that large fonts were significantly less credible than both small or medium fonts ($p < .01$), but that small and medium fonts were not detectably different.

Recall, however, that Font Size was involved in a three-way interaction with Images and Words, adding nuance to the interaction present in Figure 6. Figure 8 shows the same plot as Figure 6, but now broken out by three levels of Font Size. It is clear that as Font Size changes, the Images × Words interaction also changes. Specifically, for small and large fonts, the medium word count brings credibility ratings for all three image levels together, but for medium fonts, credibility ratings at the medium word count diverge, converging instead more at the low word count. These observations are confirmed by significant interaction contrasts ($p < .05$).

**Page time**

We also measured how long participants spent on each page while forming their credibility judgments, as page times might indicate the confidence of such judgments (Tractinsky et al., 2006). Faster times suggest that credibility judgments were easily formed, whereas slower times suggest more scrutiny.
was necessary. Also, basic sanity checking is available through an examination of page times.

An omnibus test showed that there were significant main effects of Video ($F_1, 676.5 = 9.16, p < .01$), Images ($F_2, 676.9 = 15.42, p < .0001$), and Words ($F_2, 683.9 = 10.02, p < .01$) on Page Time. There were no significant interactions among these factors. The following paragraphs discuss each effect in turn.

Figure 9 shows Page Times for when video was absent or present. The mere presence of a typically unwatched video increased time-on-page by just over one second on average.

Figure 10 shows Page Times for each level of Images. Expectedly, more images in an article resulted in more time spent, even when those images were content-free. Post hoc pairwise comparisons using Holm’s sequential Bonferroni procedure (Holm, 1979) indicate that page times were significantly different among all three levels of Images ($p < .01$). It seems each image added about 350 milliseconds, on average.
Figure 9. Average page times by video presence. Error bars and values in parentheses represent +1 SD.

Figure 10. Average page times by number of images. Error bars and values in parentheses represent +1 SD.
Finally, Figure 11 shows how Page Times were affected by the number of “lorem ipsum” words. As with videos and images, having more words resulted in longer page times. Post hoc pairwise comparisons using Holm’s sequential Bonferroni procedure (Holm, 1979) indicate that many words produced significantly longer page times than either medium or few words ($p < .01$), but that medium and few word counts were not detectably different.

Overall, then, it seems that times-on-page generally increased with increasing content (videos, images, and words). This result might seem unsurprising, but what is noteworthy here is that the content was known by participants to be devoid of any meaning. There was little, if anything, for participants to consume in the content itself. Nonetheless, participants were obligated to visually process web pages’ stylistic features in order to make a judgment about perceived credibility. And in that regard, page content—even meaningless content—seems to have played a role in forming credibility judgments.

The page time results also allow us to assess the lighthearted claim that “a picture is worth a thousand words”. Each image added about 350 ms of page time (Figure 10). Similarly, every 100 words added about 310 ms (Figure 11). Thus, it seems that, at least in the case of heavily blurred images and “lorem ipsum” words, a picture is worth about 100 words in terms of page time.

**Figure 11.** Average page times by word count. Error bars and values in parentheses represent +1 SD.
Interview results

Immediately after participants completed the study, we conducted a semi-structured interview to discover qualitative insights about their experiences. We asked participants about their first impressions, how they judged credibility, and which page elements they felt they noticed most. When appropriate, we asked follow-up questions to encourage participants to expound upon their responses.

Twelve of our 31 participants (39%) said that the presence of images or videos was most impactful on their credibility ratings, and seven participants also commented on the placement of images and videos. P11 said, “When there was a video in the middle of the article, I thought it looked less like a news article, since articles usually put the videos at the very top, since that’s the main point”.

Ten of our participants (32%) commented on font size being the factor they felt had the most impact on their credibility ratings, specifically that larger fonts lowered a page’s perceived credibility. This sentiment is in agreement with our quantitative findings about large fonts and lower credibility (Figure 7), especially for short and long articles with no images (Figure 8).

Many participants’ statements evidenced interactions between factors, such as when P6 said, “If there was only a little text, and a lot of pictures, then I would find [the article] less believable”. This comment agrees with the quantitative finding that six images were rated as less credible than three images when both were in the presence of only a few words (Figure 6). (We see this finding emerge in our second study, below, as well.)

Overall, participants were clear that having some images and videos made articles seem more credible, but that having too many images reduced credibility. However, there was a lot of interaction between factors: most participants identified the impact of certain factors as being dependent upon the levels of other factors, while some just said there needed to be a balance among levels of factors. This notion of “balance” in the visual design of newslike articles also emerged prominently in our second study.

Other comments were quite specific about individual factors. For example, P14 said she was “looking at the how big the title was—you want to catch the person’s attention, but at the same time, I feel like a lot of clickbait articles have huge titles because that’s all they focus on, so I tried to see how big that was”.

Thus, our interview results corroborated our quantitative findings. Participants’ evaluations of credibility indicated that certain factors had a bigger impact on credibility ratings than others—namely, video and image presence and placement, image count, and the interaction among image count, word count, and font size.

We now turn to our second study investigating credibility perceptions of online news articles based on their visual appearance, this time, using actual news articles from the Web.
Study 2: method

Like our first study, our second study was designed to elicit participants’ credibility perceptions of online news based on articles’ visual appearance and not their content. Rather than doing so with synthetic articles devoid of meaning, as in the first study, our second study utilised actual news articles, half from credible and half from non-credible sources, unbeknownst to participants. These articles were presented very rapidly so that their content could not be apprehended, allowing only their visual appearance to make quick first impressions (Lindgaard et al., 2006; Tractinsky et al., 2006). Our rapid exposure protocol was similar to others used in psychology for preventing deep processing of visual stimuli (Potter, 1984). By utilising this complementary method of isolating visual appearance from substantive content, we could triangulate the findings from our first study.

Participants

We recruited 30 university students for our second study, none of whom had participated in our first study. Nineteen participants were female (63%) and 11 were male (37%); the mean age was 22.6 years ($SD = 2.7$). Participants were recruited by university-wide email lists and by word-of-mouth. Each participant was compensated with a $15 USD Amazon.com gift card for about 30 min of their time.

Whereas in our first study we focused on undergraduate students, in our second study, we expanded our focus to university students generally, including graduate and professional students. Eleven such students participated. As a result, our participants’ ages ranged from 18 to 32 years. Therefore, we caution against generalising our findings beyond this young adult age range.

Of our 30 participants, 14 were from engineering or computing fields, nine were from arts or sciences, three were from business, two were from global health, and two were from architecture.

Apparatus

To conduct our study, we built a full-screen Microsoft Windows desktop application capable of presenting newslike articles gathered from the Web. In February and March 2020, we gathered 200 articles from 18 popular online sources. To ensure faithful capture and reproduction, the full-length articles were snapshotted as images. Half of the articles were from five credible news sources appearing in the top-10 most popular online news outlets on the Web. The other half of the articles were from 13 popular fake news or

---

4Our question for gender was open-ended: “With what gender do you identify, or you can prefer not to respond?” All participants chose to respond with either “female” or “male”.

satirical sources according to Snopes.com and MediaBiasFactCheck.com, which are media fact-checking and myth-debunking websites. Figure 12 shows top portions of three example articles, one each from a credible source, a questionable source, and a satirical source.

Articles were selected at random, but were required to display no overt political biases, outlandish claims, shocking images, or other content that could provoke an immediate reaction or betray it as non-credible. As a result, articles were “neutral” in their tone, imagery, and subject matter. All articles fit into one of 10 categories: science, environment, health, business, world news, local news, politics, military, law, or entertainment. Authors, affiliations, and brands were removed from all articles. Advertisements were also removed because we were interested in participants’ first impressions of the articles themselves, not of advertisements. Prior work suggests that a lack of advertisements should be unlikely to influence young adults’ perceptions of news articles (Howe & Teufel, 2014), and common technologies like ad blockers remove ads for some users anyway. Because all articles were snapshotted as images, videos did not play, but were captured while paused on their first key frame, showing their overlaying playback controls. Table 1 gives more information about our articles’ sources.

Each article was measured for various properties: word count, body font size, title font size, serif vs. sans serif font face, image count, video count, video placement, and number of embedded tweets. Font sizes were measured in pixels directly off the captured article images. These properties, and others derived from them, are discussed further below.

**Procedure**

Our study was originally designed to take place in a laboratory setting, but due to the COVID-19 pandemic, we adapted our study to take place online. We created a website\(^6\) containing all of our study materials, including our consent form and the study software. Using Zoom videoconferencing software, we guided participants in downloading the study software and running it. At the end of the study session, we asked participants a few questions about what they felt they noticed most when making their credibility judgments.

Our study software ran topmost and full-screen to prevent interruptions from other software applications. It presented 100 trials to each participant. On any given trial, the software displayed a “3-2-1” countdown followed by an article selected randomly (with replacement) from among the 200 articles in our collection. Participants had no control over the articles and no ability to interact with them. Rather, the articles were presented and automatically scrolled at the equivalent reading speed of 5600 words per minute (WPM). This rate was chosen as a value about 20 times faster than typical reading

\(^6\)http://depts.washington.edu/acelab/proj/news/.
speeds (Carver, 1992; Legge et al., 1985; Seidenberg, 2017; Witty, 1969), making it impossible for participants to read the articles. Automatic scrolling speeds were calculated based on article heights and word counts to achieve precise timing for 5600 WPM. The resultant experience for participants was one of the articles flashing by, being visible briefly but leaving only a visual first impression.

After articles were scrolled automatically to the bottom, they disappeared, and participants were prompted with an agreement-based Likert scale with the prompt, “The article I just saw seemed believable” (Figure 13). Available responses ranged from $1 = “Strongly Disagree” to $7 = “Strongly Agree”. Participants could select their rating using either the mouse or, more commonly, by pressing a key (1-7). After each rating, the scale disappeared and the “3-2-1” countdown appeared again, followed by the next article. Participants were prompted by the software for a short break every 20 trials. We remained present via Zoom videoconferencing for the entirety of the study session.

After rating the 100 articles presented to them, participants were asked to respond freely in a semi-structured interview to questions concerning what they noticed most when making their credibility ratings. As in our first study, we recognise that participants’ answers to these questions cannot be taken as ground truth.

---

7Some readers might note that rounding, quantization, and the fact that Microsoft Windows is not a real-time operating system all dictate that exact timing at 5600 WPM would be impossible to achieve in practice. To verify our timing, we logged the actual presentation time of each article for every trial, finding that the absolute mean difference between articles’ target reading speeds and presented reading speeds was 7.68 WPM (SD = 10.26), which amounted to, on average, a timing discrepancy of only 6.8 ms per article, or 0.14% timing error.
for participants’ actual behaviour, but they are useful to discover what participants thought mattered to them in forming their credibility judgments.

**Study design & analysis**

Strictly speaking, our second study was not an experiment, because participants were not randomly assigned to controlled treatment conditions. Rather, each

---

### Table 1. Sources used in our collection of 200 newslike articles gathered from the Web. All MediaBiasFactCheck URLs begin with https://mediabiasfactcheck.com where ellipses (“…”) appear.

<table>
<thead>
<tr>
<th>Source</th>
<th>#</th>
<th>URL</th>
<th>MediaBiasFactCheck Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cable News Network</td>
<td>20</td>
<td><a href="http://www.cnn.com">www.cnn.com</a></td>
<td>/cnn/ Left, mixed factual reporting</td>
</tr>
<tr>
<td>The Guardian</td>
<td>20</td>
<td><a href="http://www.theguardian.com">www.theguardian.com</a></td>
<td>/the-guardian/ Left-center, high factual reporting</td>
</tr>
<tr>
<td>India Times</td>
<td>20</td>
<td><a href="http://www.indiatimes.com">www.indiatimes.com</a></td>
<td>/india-times-bias-rating/ Right-center, mixed factual reporting</td>
</tr>
<tr>
<td>Washington Post</td>
<td>20</td>
<td><a href="http://www.washingtonpost.com">www.washingtonpost.com</a></td>
<td>/washington-post/ Left-center, high factual reporting</td>
</tr>
<tr>
<td>Boston Leader</td>
<td>5</td>
<td><a href="http://www.bostonleader.com">www.bostonleader.com</a></td>
<td>/N/A² Fake news, some satire</td>
</tr>
<tr>
<td>Breitbart</td>
<td>8</td>
<td><a href="http://www.breitbart.com">www.breitbart.com</a></td>
<td>/breitbart/ Extreme right, propaganda, conspiracy, failed fact checks</td>
</tr>
<tr>
<td>Conservative Daily</td>
<td>7</td>
<td>conservativedailypost.com</td>
<td>/conservative-daily-post/ Extreme right, propaganda, conspiracy</td>
</tr>
<tr>
<td>D.C. Gazette</td>
<td>10</td>
<td>thedcgazette.com</td>
<td>/dc-gazette/ Extreme right, conspiracy, propaganda, low factual reporting</td>
</tr>
<tr>
<td>Gateway Pundit</td>
<td>10</td>
<td><a href="http://www.thegatewaypundit.com">www.thegatewaypundit.com</a></td>
<td>/the-gateway-pundit/ Extreme right, propaganda, conspiracy, nationalism, fake news, low factual reporting</td>
</tr>
<tr>
<td>InfoWars</td>
<td>10</td>
<td><a href="http://www.infowars.com/news/">www.infowars.com/news/</a></td>
<td>/infowars-alex-jones/ Conspiracy, pseudoscience, very low factual reporting</td>
</tr>
<tr>
<td>Prntly</td>
<td>5</td>
<td>prntly.com</td>
<td>/prntly/ Extreme right, propaganda, fake news, low factual reporting</td>
</tr>
<tr>
<td>Burrard Street Journal</td>
<td>10</td>
<td>burrardstreetjournal.com</td>
<td>/burrard-street-journal/ Satire</td>
</tr>
<tr>
<td>National Report</td>
<td>5</td>
<td>nationalreport.net</td>
<td>/national-report/ Satire</td>
</tr>
<tr>
<td>Now 8 News</td>
<td>5</td>
<td>now8news.com</td>
<td>/now8news/ Satire</td>
</tr>
<tr>
<td>The Onion</td>
<td>5</td>
<td><a href="http://www.theonion.com">www.theonion.com</a></td>
<td>/the-onion/ Satire</td>
</tr>
<tr>
<td>Real News Right Now</td>
<td>10</td>
<td>realearnsrightnow.com</td>
<td>/real-news-right-now/ Satire</td>
</tr>
<tr>
<td>World News Daily</td>
<td>10</td>
<td>worldnewsdailysreport.com</td>
<td>/world-news-daily/ Satire</td>
</tr>
</tbody>
</table>

²Boston Leader comes from Associated Media Coverage and is mentioned on Snopes.com’s 2016 review of fake news websites: https://www.snopes.com/news/2016/01/14/fake-news-sites/.

---

![Image](image.png)

**Figure 13.** The 7-point Likert scale used to capture credibility ratings in our second study.
participant was exposed to 100 randomly selected articles, each with various properties, and it is the correlation of those properties with participants’ credibility ratings that we are interested in. Therefore, this study can make no claim to what causes participants to perceive an article as credible or not, but it can shed light on the visual properties of articles that correlate with news credibility ratings.

The following article properties constituted the predictors (scalar variables) and factors (categorical variables) that could correlate with participants’ credibility ratings. Means ($M$), standard deviations ($SD$), and ranges are for all analysed trials in the study.

- **Words**: scalar; $M = 684$, $SD = 509$, range = [117, 2995]
- **Body Font Size**: scalar; $M = 13.1$ px, $SD = 1.5$ px, range = [9, 16]
- **Title Font Size**: scalar; $M = 31.9$ px, $SD = 7.0$ px, range = [15, 48]
- **Font Face**: categorical; levels = { serif, sans serif }
- **Images**: scalar; $M = 1.88$, $SD = 1.57$, range = [0, 11]
- **Videos**: scalar; $M = 0.30$, $SD = 0.61$, range = [0, 4]
- **Video Placement**: categorical; levels = { top, middle }
- **Tweets**: scalar; $M = 0.23$, $SD = 0.70$, range = [0, 8]

In addition, during the variable selection process (Heinze et al., 2018), the following computed properties were considered out of a realisation that longer articles have more room for images, videos, and embedded tweets, and so the “per word density” of these elements might explain more variance in the responses than their raw counts.

- **Image Density**: scalar; $M = 0.00350$, $SD = 0.00275$, range = [0.00000, 0.01741]
- **Video Density**: scalar; $M = 0.00058$, $SD = 0.00136$, range = [0.00000, 0.00855]
- **Tweet Density**: scalar; $M = 0.00046$, $SD = 0.00136$, range = [0.00000, 0.00917]
- **Has Video**: categorical; levels = { yes, no }  
- **Has Tweet**: categorical; levels = { yes, no }

In addition, as in the first study, we examined participants’ Age and Gender.

Our two dependent variables were participants’ Credibility ratings, an ordinal measure, and Rating Time, a scalar measure in milliseconds. Unlike the first study, where participants controlled how much time they spent on each article, in this second study, article exposure time was controlled by the software. However, we measured how long each participant took to provide their credibility rating on the 1–7 scale after an article appeared, was autoscrolled, and disappeared, which might indicate the level of confidence a participant has in their rating (Tractinsky et al., 2006).

With 30 participants and 100 trials per participant, we collected 3000 total trials. However, one participant reported feeling ill after trial 60, marking the
same “Neutral” (4) response on the Likert scale repeatedly thereafter. Therefore, her last 40 trials were discarded. Furthermore, of the remaining 2960 trials, 16 showed rating times longer than 10 seconds, which is an unrealistically long amount of time and likely indicates an outside interruption or ill-timed hiatus. These trials were also discarded, resulting in 2944 trials that constituted our final data set.

To build the best-fitting model relating our predictors and factors to our dependent variables, we utilised typical variable selection approaches: univariate selection, forward selection, and backwards elimination (Heinze et al., 2018). The latter two procedures used Bayesian information criterion (Schwarz, 1978) for term selection to promote model parsimony. For Credibility, these procedures converged to include Words, Body Font Size, Font Face, and Image Density as model terms. For Rating Time, only Words and Image Density emerged as model terms. In both models, Trial and Subject were included as random effects to account for repeated measures (Frederick, 1999; Littell et al., 1998).

As Credibility is an ordinal response, it was analysed with mixed ordinal logistic regression (Hedeker & Gibbons, 1994; Mangiafico, 2016). Unsurprisingly, Rating Time was found to be lognormally distributed (Lawrence, 1988; Limpert et al., 2001) and therefore the logarithm of Rating Time was analysed using a linear mixed model analysis of variance (Frederick, 1999; Littell et al., 1998). Interactions were limited to second order for interpretability. Statistical tests were conducted in R using the ordinal, RVAideMemoire, lme4, and car packages.

**Study 2: results**

In this section, we present the results of our second study of online news credibility perceptions based on presentational factors. We first discuss Credibility and then Rating Time. Then we share our interview results.

**Perceived credibility**

Recall that participants each rated 100 newslike articles on 1–7 Likert scales ranging from 1 = “Strongly Disagree” to 7 = “Strongly Agree” in response to the prompt, “The article I just saw seemed believable”.

An omnibus test shows that there were significant main effects of Words ($\chi^2_{(1,N=2944)} = 3.97, p < .05$) and Body Font Size ($\chi^2_{(1,N=2944)} = 33.65, p < .0001$) on Credibility. There was also a marginal effect of Image Density ($\chi^2_{(1,N=2944)} = 3.07, p = .080$). By contrast, Font Face did not exert a detectable main effect ($\chi^2_{(1,N=2944)} = 1.98, n.s.$). However, there was a significant Words × Body Font Size interaction ($\chi^2_{(1,N=2944)} = 3.96, p < .05$), a significant Words × Font Face interaction ($\chi^2_{(1,N=2944)} = 3.86, p < .05$), and a
significant $Words \times Image\ Density$ interaction ($\chi^2_{(1,N=2944)} = 40.28, p < .0001$), indicating that the effects of font size, font face, and images all depend on article length. In the following paragraphs, we unpack each of these results.

Figure 14 shows Credibility ratings for articles of increasing word counts. Incidentally, the average word count of presented articles was 684 ($SD = 509$), which is quite close to the medium-length articles from the first study (644). A clear upward trend is visible where longer articles tended to be perceived as more credible, at least until the very longest articles, where credibility dropped somewhat.

Unsurprisingly, body font size was correlated with title font size (Pearson $r = 0.207, N = 2944$). Given the former’s inclusion in the model, the latter was not selected as it did not explain sufficient additional variance in the response. Figure 15 shows how Credibility ratings increased with larger body font sizes.

That increasing font sizes increased perceived credibility might seem to contradict the finding about font size from the first study, where the largest fonts were least credible. However, closer inspection reveals the findings from both studies to be consistent. Font sizes in the first study were generated in points, whereas in the second study, they were measured directly off the snapshotted articles in pixels. Table 2 indicates how the two sets of font sizes relate.

In the first study, medium-sized fonts had the highest average credibility (see Figure 7). For serif fonts, the first study’s medium body font size equates to 14 px in the second study. For sans serif fonts, the first study’s medium body
font size equates to 15 px in the second study. Figure 16 shows that for each font face, Credibility was high around these values in the second study, too, but then seems to decrease with the largest fonts. In the second study, we did not have any articles with body fonts as large as the largest fonts from the first study (17 and 18 px for serif and sans serif fonts, respectively). Thus, it seems medium-sized fonts were most credible in both studies.

During variable selection (Heinze et al., 2018), it became clear that Images explained some variance in Credibility, but Image Density—the ratio of images to words—was a better predictor in the model, which is intuitive because longer articles have more room for images before appearing cluttered. (Videos and tweets were not numerous enough in our articles to exert the same effect.) Although Image Density only exhibited a marginal effect on Credibility ($p = .080$), Figure 17 hints at a downward trend—as image density increased, perceived credibility decreased.

### Table 2. Font sizes in points (pt) and their equivalent sizes in pixels (px).

<table>
<thead>
<tr>
<th>Font Face</th>
<th>Font Size</th>
<th>Body (pt)</th>
<th>Title (pt)</th>
<th>Body (px)</th>
<th>Title (px)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serif</td>
<td>Small</td>
<td>13</td>
<td>30</td>
<td>11</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>16</td>
<td>38</td>
<td>14</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>19</td>
<td>46</td>
<td>17</td>
<td>41</td>
</tr>
<tr>
<td>Sans serif</td>
<td>Small</td>
<td>13</td>
<td>30</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>16</td>
<td>38</td>
<td>15</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>19</td>
<td>46</td>
<td>18</td>
<td>44</td>
</tr>
</tbody>
</table>
Although *Image Density* provided for a better statistical model than *Images*, an examination of the latter in the second study showed an effect on perceived credibility similar to that in the first study (see Figure 5). Figure 18 indicates that, on average, few (0-1) images were least credible, but medium (2-4) and many (5-7) images were more credible. *Post hoc* pairwise comparisons with the correction from Tukey (1949) confirm these differences ($p < .01$).

As reported above, word count was involved in three two-way interactions; that is, the effects of *Body Font Size*, *Font Face*, and *Image Density* on *Credibility* varied based on article length. Figure 19 shows these interactions. As word count increased, perceived credibility increased (a) steadily for medium-sized fonts, but less so for small and large fonts; (b) more for serif fonts than sans serif fonts; and (c) most for articles with middle levels of image density.

**Distinguishing credible from non-credible articles**

The articles used in the second study were from 18 different news sources on the Web (Table 1). Five of these sources were devoted to medium-to-high levels of fact-based reporting and can be considered “credible sources”. Seven of these sources engage in low levels of fact-based reporting, push extreme right-wing agendas, and promulgate conspiracy theories, pseudoscience, fake news, and propaganda; these are considered “questionable sources”. Finally,
Figure 17. Average credibility ratings by images per word. Note that no articles in our collection had between 0.014 and 0.016 images per word, resulting in the discontinuity. Higher is “more credible”. Error bars represent ±1 SD.

Figure 18. Average credibility ratings by image count. Higher is “more credible”. Error bars and values in parentheses represent +1 SD. The effect of images on perceived credibility is similar to that in the first study, whose image counts were 0, 3, and 6.
six sources were devoted to satire, which, like questionable sources, constitute fake news, but with an attempt at humour.

As described above, our participants saw each article only briefly, with articles appearing only for a duration equal to a reading speed of 5600 words per minute. Participants could not read the articles or scrutinise their images. So, it is interesting to consider whether a quick first impression of the articles provided enough of a signal for participants to distinguish between credible and non-credible sources.

Figure 20 shows average Credibility for each Source Type. Mixed ordinal logistic regression (Hedeker & Gibbons, 1994; Maniafico, 2016) confirms the ratings were significantly different for each type of source ($\chi^2(2, N=2944) = 375.92, p < .0001$). Post hoc pairwise comparisons with the correction from Tukey (1949) indicate that all levels of Source Type were detectably different ($p < .0001$). Thus, it seems that even with only brief exposure, participants were able to detect a difference between credible and non-credible sources. Across our two
studies, we see clues as to which visual properties seem to be informing participants’ judgments. We take up this point further in our discussion, below.

An examination of credibility ratings for each of the 18 article sources shows that not all fact-based sources were rated as most credible, and some questionable and satirical sources were rated as appearing more credible than some fact-based sources (Figure 21).

**Rating time**

Recall that in our first study, participants spent as much time on each synthetic “article” as they liked. In this second study, however, article presentation was fully controlled by the software. After each article was scrolled rapidly, it disappeared, and the participant was prompted to make a credibility rating (see Figure 13). How long participants took to provide this rating can indicate how confident they were of their judgments (Tractinsky et al., 2006). An omnibus test shows that there was a significant main effect of *Image Density* on *Rating Time* ($F_{1, 2875.8} = 22.65$, $p < .0001$). There was also a significant *Words* × *Image Density* interaction ($F_{1, 2876.2} = 12.84$, $p < .0001$). These results are explained below.

Figure 22 shows rating times for different article lengths and image densities. A “U”-shaped pattern is visible, where shorter and longer articles took somewhat more time to rate than medium-length articles. This “U”-shape is least
pronounced for articles with the fewest images. Short articles with many images were slow to rate, as were long articles with few images. If fast ratings are taken to suggest confidence in credibility judgments (Tractinsky et al., 2006), then participants were most confident when rating medium-length articles containing a medium number of images.

Above, we reported that participants rated fact-based articles as more credible than articles from questionable or satirical sources (Figure 20). If rating time is taken to be suggestive of rater confidence (Tractinsky et al., 2006), then we can also say that participants seemed more confident when rating articles from credible sources than articles from non-credible sources (Figure 23). There was a significant effect of Source Type on Rating Time ($F_{2, 2875.7} = 15.62, p < .0001$). Post hoc pairwise comparisons with the correction from Tukey (1949) indicate that fact-based rating times were significantly lower than those for questionable or satirical articles ($p < .001$), but rating times for questionable and satirical articles were not detectably different.

**Interview results**

After participants completed their study sessions, we sought to learn what they thought mattered most when forming their credibility judgments. We asked them, “What did you notice most when making your article believability ratings?” We did not ask leading questions but allowed participants to guide
Figure 22. Average rating times, in milliseconds, by word count and images per word. Error bars are omitted for readability. Lower values are faster and can indicate higher rater confidence (Tractinsky et al., 2006).

Figure 23. Average rating times, in milliseconds, by the source type for each article (see Table 1). Error bars and values in parentheses represent +1 SD. Lower values are faster and can indicate higher rater confidence (Tractinsky et al., 2006).
the discussion. Participants’ comments both validated our study methodology and corroborated our quantitative findings.

In total, participants made 169 comments referring to properties of news articles affecting their credibility judgments. Of these, 142 comments, or 84%, referred only to visual properties of articles, suggesting that our rapid exposure protocol was effective in separating articles’ appearance from content. The other 27 comments reflected only surface features of articles’ contents—e.g. a particular word in a title (P22), a political figure in a photo (P15), or that an article seemed generally about science instead of politics (P27). Such comments indicated that some shallow content was still apprehensible to participants even at a reading speed of 5600 words per minute, but such comments were relatively few (16%) and non-specific.

Article length, i.e. word count, was mentioned by 12 of 30 participants (40%). Specifically, four participants said long articles increased credibility, and one participant said long articles decreased it. Moreover, seven participants said short articles decreased credibility. These sentiments were consistent with Figure 14, showing an inverted “U” shape for credibility ratings by word count.

Not surprisingly, images were a salient visual feature for participants. Although articles were presented too quickly for participants to scrutinise their images, participants still could grasp images’ gist. Eleven of 30 participants (37%) said that high quality, professional looking photographs increased credibility. Twelve of 30 participants (40%) said that serious newslike images increased credibility compared to sensational or humorous images. As P4 said, “Paparazzi-style photos looked less credible”.

Interestingly, without being prompted, eight of 30 participants (27%) made comments indicating that higher credibility was achieved by balancing article length and image count, i.e. Image Density, and that a medium number of images was seen as most credible (Figures 18 and 19(c)). As P12 said, “Having an even ratio of pictures to text made an article seem more credible”. Similarly, P26 said, “[It] is more credible when there is a balance between photos and text”. And P30 said, “If there are too many images compared to text, [articles] seemed less believable”. It is clear that participants were cognizant of the ratio of images to text in their credibility judgments.

Sixteen of 30 participants (53%) mentioned that serif fonts increased perceived credibility compared to sans serif fonts; no participants said the opposite. As P2 said, “Times New Roman and other serif fonts were most believable”. Referring to fonts, P16 said, “Formats that resemble traditional [print] newspapers were more credible”. That said, recall that Font Face alone did not exert a detectable effect on Credibility, but a Words × Font Face interaction did (Figure 19(b)).

Only four of 30 participants (13%) mentioned font size, with two participants indicating that large fonts decreased credibility, and two indicating that
small fonts did so. That large and small fonts were mentioned for decreasing credibility, but medium-sized fonts were not mentioned, is consistent with quantitative findings from both the first and second studies about medium font sizes being most credible (Figures 7, 15, and 16; Table 2).

Participants were divided on how embedded tweets in articles affected credibility. Eleven of 30 participants (37%) mentioned that embedded tweets affected their credibility perceptions, but four participants said tweets increased credibility while seven participants said they decreased it. A few participants said that tweets appearing in science, environment, health, or business articles decreased credibility; but tweets appearing in politics or entertainment articles increased credibility. These results explain why neither Tweets nor Tweet Density had significant effects on Credibility. Of course, these perceptions are despite participants not knowing what any tweets actually said.

Ten of 30 participants (33%) mentioned that the presence of hyperlinks affected their credibility judgments, with nine indicating that hyperlinks increased perceived credibility and one indicating they decreased it. Unlike in the first study, hyperlink densities were not extracted from articles in the second study; this question remains open for future work.

**Discussion**

Having completed two complementary studies that isolated the visual appearance of newslike articles in entirely different ways, we can see certain common findings emerge. In both studies, articles with a medium number of images, about three to seven, were more credible than articles having very few or very many images (Figures 5, 6, 8, and 18). It is also clear that the images-to-text ratio mattered; too many or too few images relative to article length lowered credibility (Figures 6, 8, 17, and 19(c)). In both studies, font size mattered, with 14–16 pixel body fonts emerging as most credible, especially for serif fonts (Figures 7, 15, 16, and 19(a); Table 2). It is also clear from both studies that article length interacts with these factors (Figures 6, 8, and 19).

Although our two studies converged on multiple findings, there were also a few differences in their results. In the first study, video presence had a significant effect on perceived credibility (Figure 4), which did not emerge in the second study. That said, videos in the second study were not interactive, and were more akin to images, which showed effects in both studies. Also, in the first study, article length did not, by itself, have a detectable effect on perceived credibility, but in the second study, it did (Figure 14), with credibility peaking at 2250 words, long enough for the images per word ratio to be low enough to increase credibility (Figure 19(c)).

Taken together, the two studies suggest that when it comes to the visual properties of online news, credibility ratings are highest when articles are in
a kind of “Goldilocks zone”, achieving a “just right” balance among certain properties. For example, increasing word count increases credibility, but only to a point (Figure 14). Similarly, increasing font size increases credibility, but only to a point (Figures 7, 15, and 16; Table 2). Credibility seems highest with a medium number of images (Figures 5 and 18), provided that they are not too dense relative to word count (Figures 6 and 17). Using rating time as a proxy for rater confidence (Tractinsky et al., 2006), participants seemed most confident in making credibility judgments of medium-length articles containing a medium-number of images (Figure 22). And participants’ interview responses in both studies indicate that they were cognizant of “balance” in article appearance, whether it be for font size, word count, image count, or the ratio of images to text. The cliché that “the total is more than the sum of its parts” seems at work here, where the overall gestalt of a page is responsible for participants’ credibility judgments more than any single presentational factor.

Participants seemed adept at distinguishing credible from non-credible news articles in the second study, even at reading speeds of 5600 words per minute (Figure 20). Even still, as Figure 21 shows, some non-credible sources were perceived as more credible than some credible sources. The lowest-rated fact-based source was the India Times ($M = 4.08, SD = 1.24$), while the highest-rated questionable source was Breitbart ($M = 4.49, SD = 1.43$). The highest-rated satirical source was Real News Right Now ($M = 4.46, SD = 1.37$). Figure 24 shows a representative article from each of these sources. At a glance, it is not easy to tell which article is from the most credible source.

**Design implications**

Taken together, our two studies have multiple implications for the design of online news. For example, the least credible article would exhibit extremes: it would either be very short (<400 words) or very long (>2500 words) (Figure 14); it would contain either very few (0-1) or very many (>8) images, depending on its length (Figures 5, 6, 17, 18, and 19(c)); and it would employ either very small (11/28 px) or very large (18/44 px) body/title sans serif fonts (Figures 7, 8, 15, 16, and 19(a,b); Table 2).

The most credible article would avoid these extremes and would appear balanced: it would be long, but not too long (1100-2250 words) (Figures 6, 8, and 14); it would have a few (3-7) but not too many images, depending on article length (Figures 5, 6, 17, and 18), and possibly a video (Figure 4); and it would employ medium (13/27 px – 15/36 px) body/title serif fonts (Figures 7, 8, 15, 16, and 19(a,b); Table 2). Of course, the exact values of these settings might vary from those we tested and observed, but the general direction of these findings should hold. Figure 25 shows top portions of three article layouts, two that exemplify “least credible” designs and one “most credible” design.
A concerning set of design implications pertains to people creating online news sites to deliberately propagate falsehoods. Those trying to increase the perceived credibility of their site’s articles could conceivably make use of our findings to make those articles look more believable. However, the same could be said for legitimate online news sources seeking to reinforce their own credibility. In fact, if all online news sources converged upon the Goldilocks zone to maximise their apparent credibility, then it would effectively remove visual appearance as an influence, creating a de facto standard whereby credibility is based mostly on article content. Furthermore, by uncovering some of the purely presentational factors that affect credibility, citizens...
can become more aware of their perceptions and “look beneath the surface” to carefully scrutinise the credibility of their news sources.

**Limitations of the studies**

As with any studies, ours had limitations. Numerous other presentational factors could have been examined in either study but were deemed out of scope. Some of these other factors were line and margin spacing, text justification, colours, fonts, image and video sizes, image placements, and advertisements. The Web is a rich environment—isolating precisely which elements affect credibility, and how, is an ambitious undertaking, and our two studies only begin to shed light on the interplay of the many factors at work.

Another limitation of our studies was that 61 university-aged students (18–32 years old) were our combined participant pool. As mentioned, this demographic ranks high among social media use and online news promulgation and consumption (Pew Research, 2019). But our study findings might not generalise beyond young adults. A fuller picture could be obtained by including other and more diverse participants, participants of different ages, different educational backgrounds, members of various political parties, and people from different geographies.

Another limitation is that our newslike articles in both studies were viewed on desktop or laptop computers, not tablets or smartphones. Many websites employ “responsive” designs that change their layouts to accommodate different form factors. And many young adults get their news on small screens like those of smartphones. Whether our findings would generalise to these small screens remains an open question.

Whereas our first study generated synthetic articles with predetermined properties, our second study utilised actual news articles sampled from popular outlets; the articles therefore reflected the natural variation of properties in the sources from which they were drawn. Therefore, the factors correlated with perceived credibility cannot be said to cause changes in perceived credibility; our study is only able to establish correlations and associations, not causation. To establish the latter, participants would need to be randomly assigned to conditions reflecting controlled variations in the factors of interest, and hold other factors constant. Our first study assigned participants to controlled treatments, but our second study did not.

Our two studies were focused on isolating, inasmuch as possible, presentation from content to examine the former’s effects on perceived credibility. Therefore, our findings cannot establish how presentation interacts with content. Whether our findings would hold when participants have time to absorb both presentation and content is an open question. In our first study, participants had unlimited time with each article, but no article content. In our second study, participants had article content but very little time. Neither study can shed
light on which factors affect credibility when participants have both time and content. In terms of Fogg’s (2003) *Prominence-Interpretation Theory*, we have examined what visual aspects of articles make them prominent, which is a prerequisite for judging articles as credible; but such credibility judgments can fully emerge only after users interpret what they see and form judgments.

**Future work**

Beyond addressing the limitations raised above, future work on this topic is replete with interesting directions. With the knowledge gleaned from these studies, we can begin to understand how presentational factors promote or diminish perceived credibility, independent of content.

A promising future project could be to apply systematic distortions (Spillane et al., 2017) to news articles with actual content—perhaps using a web browser plugin—and then observe whether the distortions we apply do indeed correlate with changes in credibility perceptions in ways predicted by our findings. For example, if a browser plugin were able to change the font faces and font sizes of online news articles, would we see credibility perceptions change accordingly?

A limitation of our studies was not knowing exactly what participants were looking at, and for how long. An eye-tracker would provide this information to accompany credibility ratings and participants’ interview responses. Understanding how objective eye-tracking measurements might correspond with our findings would be valuable.

We can imagine creating a “credibility rater” program as a web browser plugin that provides users with an assessment of the credibility of an article based on the many visual features considered in our studies. A deep learning model could be trained with crowdsourced inputs pairing actual articles with crowd workers’ credibility ratings. The plugin could refine its training by enabling users to disagree with the provided rating and supply the correct one. Over time, the model could account for the complex interplay of many visual features, and ultimately provide better signals to users about the credibility of news they encounter on the Web.

Lastly, our studies were limited to desktop and laptop computing environments, but many people today consume their news on smartphone or tablet devices. Replicating our study on such devices would yield an understanding of how device-independent these presentational factors are (or are not).

**Conclusion**

In this work, we presented results from two complementary studies investigating how presentational factors affect the perceived credibility of newslike
articles for young adults aged 18–32 years. Unlike prior work on web credibility, which has focused on web content, or which has blended content with presentation, our work isolated the presentation of online news in two different ways, by using: (1) synthetic “articles” generated without any meaningful content, and (2) a rapid exposure protocol where actual news articles were presented at a reading speed of 5600 words per minute, or about 20 times typical reading speeds. Our findings indicate that presentational factors, even isolated from content, do affect perceived credibility. Across the two studies, main effects and interactions emerged involving article length, image count, image density, video presence, font face, and font size. The notion of a “Goldilocks zone” seemed operative, where medium levels of factors and an overall balance of factors combined to yield the highest perceived credibility. This zone was evident not just in our quantitative results but also in our post-session interviews with participants from both studies, where “balance” in visual design was mentioned multiple times.

Our work has shed light on the presentational factors of online news that inform people’s credibility judgments, which are clearly based not only on article content but also on article appearance. It is our hope that these findings can help inform both people and systems when making credibility judgments of online news. Ultimately, we hope this work leads to more informed and better equipped consumers and promulgators of online information.

**Acknowledgements**

The authors thank Mingyuan “Jason” Zhong, Mingrui “Ray” Zhang, Alexis Hiniker, and Meredith Ringel Morris. This work was supported in part by the University of Washington Information School, the Mani Charitable Foundation, and the University of Washington Center for an Informed Public. Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect those of any supporter.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

This work was supported by the University of Washington Center for an Informed Public; Mani Charitable Foundation; University of Washington Information School.

**References**


