Revealing complexities when adult readers engage in the credibility evaluation of social media posts

Miikka Kuutila[a,b,*], Carita Kiili[c], Reijo Kupiainen[c,d], Eetu Huusko[b], Junhao Li[b], Simo Hosio[b,e], Mika Mäntylä[b,f], Julie Coiro[g] and Kristian Kiili[c]

[a]Faculty of Computer Science, Dalhousie University, 6050 University Avenue, PO BOX 15000, Halifax, N S, B3H 4R2, Canada
[b]Faculty of Information Technology and Electrical Engineering, University of Oulu, P.O. Box 4500, Oulu, 90014, Finland
[b]Faculty of Education and Culture, Tampere University, P.O. Box 700, Tampere, 33014, Finland
[d]Department of Education and Lifelong Learning, Faculty of Social and Educational Sciences, Norwegian University of Science and Technology, P.O. Box 8900, Torgarden NO-7491, Trondheim, Norway
[e]Tokyo College, University of Tokyo, 1 Chome-1-1, Bunkyo City, 113-0032, Japan
[f]Department of Computer Science, University of Helsinki, PL 68, Helsinki, 00014, Finland
[g]Feinstein College of Education, University of Rhode Island, 615 Chafee Hall, 142 Flagg Road, Kingston, RI 02881, United States

ARTICLE INFO

Keywords:
Sourcing
Trustworthiness
Social media
Prior beliefs
Crowdsourcing
Digital reading

ABSTRACT

The internet, including social networking sites, has become a major source of health information for laypersons. Yet, the internet has also become a platform for spreading misinformation that challenges adults' ability to critically evaluate the credibility of health messages. To better understand the factors affecting credibility judgements, the present study investigates the role of source characteristics, evidence quality, crowdsourcing platform, and prior beliefs of the topic in adult readers' credibility evaluations of short health-related social media posts. Researchers designed content for the posts concerning five health topics by manipulating source characteristics (source's expertise, gender, and ethnicity), accuracy of the claims, and evidence quality (research evidence, testimony, consensus, and personal experience) in the posts. Then, accurate and inaccurate posts varying in these other manipulated aspects were computer-generated. Crowdworkers (N = 844) recruited from two platforms were asked to evaluate the credibility of ten social media posts, resulting in 8380 evaluations. Before credibility evaluation, participants' prior beliefs on the topics of the posts were assessed. Results showed that prior belief consistency and source expertise most affected the perceived credibility of accurate and inaccurate social media posts after controlling for the topic of the post. In contrast, the quality of evidence supporting the health claim mattered relatively little. In addition, the data collection platform had a notable impact, such that posts containing inaccurate claims were much more likely to be rated higher on one platform compared to the other. Implications for credibility evaluation theory and research are discussed.

1. Introduction

The internet has become a major source of health information for laypersons (Sun et al., 2019). Because information online can be published without adhering to verification standards (Braasch et al., 2014), the internet has also become a platform for spreading misinformation, including issues concerning health (Ecker et al., 2022). This requires a level of criticality that can challenge laypersons, especially when their prior knowledge about the issue at hand is limited (Andreassen & Strømsø, 2012; Braasch et al., 2014).

The challenge of credibility evaluation is even more pronounced when people encounter health-related issues on social networking sites. Social media messages are often very short and informal and may use improper syntax or spelling (Addawood & Bashir, 2016). On social networking sites, such as Twitter, short texts can contain arguments with accurate, inaccurate, inappropriate, or missing justifications, especially when these texts involve controversial topics. Although modern societies are portrayed by the division of cognitive labor wherein individuals defer to relevant experts to ground their understanding of less familiar topics (Scharrer et al., 2017), on social networking sites anyone can share their knowledge, views, or opinions, regardless of their expertise.

*Corresponding author
ORCID(s): 0000-0002-3695-7280 (M. Kuutila); 0000-0001-9189-4094 (C. Kiili); 0000-0003-2610-2294 (R. Kupiainen); 0000-0002-8882-8501 (E. Huusko); 0000-0002-0982-0158 (J. Li); 0000-0002-9609-0965 (S. Hosio); 0000-0002-2841-5879 (M. Mäntylä); 0000-0002-2578-3748 (J. Coiro); 0000-0003-2838-6892 (K. Kiili)

Recent frameworks for credibility evaluation (Barzilai et al., 2020; Forzani et al., 2022) have depicted factors that readers can attend to when evaluating online information, including the quality of reasoning, consistency of content with one's prior knowledge or other relevant resources, the author's expertise and intentions, and text genre. Previous studies have mainly focused on a limited number of credibility evaluation aspects or employed experimental designs in which participants are assigned to specific conditions varying in credibility features. For example, S. Chinn and Weeks (2021) assigned participants to four conditions that differed in expertise (expert vs. layperson) and evidence type (statistical evidence vs. testimony) and then asked them to evaluate statements presented in the news articles. This study revealed an interaction between source expertise and evidence type such that the quality of evidence affected the perceived credibility of the laypersons but not that of the experts.

However, studies designed to only manipulate features of content and source cannot reveal all the complexities involved in credibility evaluation. Although there is substantial evidence that indicates variations in readers, texts, activities, and contexts can influence performance when making sense of information on the Internet (Coiro, 2021), few studies have investigated a more complex combination of factors (i.e., content, source, and context) that may predict how individuals evaluate the quality of social media texts and corresponding sources. Further, drawing research-based conclusions from crowdsourcing platforms whose membership represents mostly Western populations can lead to an imbalance of cultural and social perspectives (Brandt, 2022). Consequently, sampling across platforms with more diverse membership can help examine whether the findings gleaned from Western populations hold in other regions around the globe.

Therefore, the present study approached credibility evaluation more holistically by considering multiple factors, including the content (prior-belief consistency, evidence type), the source (expertise, gender, ethnicity) and the context (two crowdsourcing platforms representing multiple nationalities) across four health topics when adult readers evaluated either accurate or inaccurate social media posts. By choosing this approach, the study provides insights about the most dominating factors to consider when advancing theoretical frameworks and models of credibility evaluation. Moreover, this study advances the understanding of credibility judgments among the adult population as readers interact with information in globally networked social media contexts.

2. Theoretical Model for First- and Second-Hand Evaluation

Our study was primarily informed by a bidirectional model of first- and second-hand evaluation strategies (Barzilai et al., 2020) while also considering one additional feature of the Critical Online Resource Evaluation (CORE) framework (Forzani et al., 2022). The bidirectional model of first- and second-hand evaluation strategies is based on the content-source-integration model (Stadtler & Bromme, 2014) that suggests that readers employ two levels of strategies to resolve scientific conflicts. Readers can evaluate information validity to determine whether the presented claims are accurate (i.e., using first-hand evaluation). However, readers do not always have sufficient knowledge to make the appropriate judgment. In this case, they can resolve the conflict by turning to second-hand evaluation to determine whether the source of information is trustworthy.

In their bidirectional model of first- and second-hand evaluation strategies, (Barzilai et al., 2020) specify three first-hand evaluation strategies to judge the validity of scientific claims (i.e., distinguish accurate claims from inaccurate ones). These strategies are knowledge-based validation, discourse-based validation, and corroboration. When using knowledge-based validation, readers rely on their prior knowledge and beliefs about the topic, whereas when using discourse-based validation, readers use various discourse features, such as argument structure and quality to make judgements about the credibility of content. Corroboration, a third strategy for validating content, is when readers compare text content to what other resources state about a given issue.

The term second-hand evaluation (Barzilai et al., 2020; Stadtler & Bromme, 2014) refers to sourcing strategies that readers employ to judge a source's trustworthiness. Readers can make inferences and evaluate a source's expertise, benevolence, and integrity (Hendriks et al., 2015) by relying on available information about the source. According to the model, readers use first- and second-hand evaluation strategies reciprocally. That is, content quality judgments inform source trustworthiness judgments and vice versa. In addition to the first and second-hand evaluation strategies, readers can also evaluate context as part of the credibility evaluation process, which Forzani et al. (2022) refers to as tertiary evaluation. According to Forzani et al. (2022), context refers to the temporal, social, and political setting of the text (e.g., social media texts read in specific crowdsourcing platforms). In this study, we considered the effects of platform context, while also incorporating a slightly broader interpretation of context by considering contextual attributes of the readers evaluating the texts (i.e., adult readers representing different nationalities).

2.1. First-Hand Evaluation

In this study, we focused on credibility judgments of health-related social media posts to determine the extent to which readers rely on their prior beliefs (i.e., knowledge-based validation) and information presented in the posts (i.e., discourse-based validation) and sourcing. Next, we define relevant concepts and outline previous research associated with first- and second-hand evaluation; first-hand evaluation involves knowledge-based and discourse-based validation while second-hand evaluation focuses on sourcing. We also present research suggesting contextual attributes, such as a reader's cultural background, can affect adults' credibility evaluations in social media contexts.

2.1.1. Knowledge-Based Validation

Prior knowledge and beliefs are internal resources that readers can use to quickly evaluate the accuracy of information (Richter & Maier, 2017). When readers evaluate information using their prior beliefs, they seem to favor belief-consistent information over belief-inconsistent information (Abendroth & Richter, 2021; McCrudden & Barnes, 2016; van Strien et al., 2016). As such, readers attend to evidence that is consistent with their beliefs and resist evidence that contradicts their beliefs. This text-belief consistency effect is similar to confirmation bias (Karimi & Richter, 2023) - the term used in the psychological literature to refer to seeking or interpreting evidence in accordance with a person's existing beliefs, expectations, or hypotheses (Nickerson, 1998). However, a confirmation bias occurs when a reader starts with a particular view of a particular issue and then actively searches for additional information that upholds that view. In contrast, the text-belief consistency effect occurs during the comprehension process (Karimi & Richter, 2023), when a reader encounters information and, often unconsciously, attends to text that is consistent with their beliefs while also passing by text that contradicts their beliefs.

The text-belief consistency effect in credibility evaluation has been shown in different contexts, including isolated arguments with no indicators of context (McCrudden & Barnes, 2016), and arguments presented in the context of social media posts (Wertgen & Richter, 2023). For example, McCrudden and Barnes (2016) asked 72 high school students to evaluate eight isolated arguments about climate change. Four arguments claimed that humans affect climate change, and four other arguments claimed the opposite. Half of the arguments were strong, and half were weak. It was found that students evaluated belief-consistent arguments more favorably than belief-inconsistent arguments. However, arguments were not evaluated solely based on their belief-consistency. Namely, students evaluated strong arguments more favorably than weak arguments. Importantly, further analysis showed that students' evaluations of strong belief-inconsistent arguments and weak belief-consistent arguments did not differ.

In a second study, Wertgen and Richter (2023) examined the text-belief consistency effect when asking university students to evaluate short messages, resembling Twitter posts, on four socio-scientific controversial topics (see Experiment 1). For each topic, the messages represented opposing positions. In addition to the position, Wertgen and Richter (2023) also manipulated the source-message consistency. In the consistent source-message condition, the source (e.g., World Wildlife Fund) presented the expected position (humans cause climate change). In the inconsistent source-message condition, the source presented an unexpected position (natural causes for climate change). Participants read belief-consistent messages faster and judged them more plausible than belief-inconsistent messages. The same pattern was found for source-message consistency. When the sources presented expected positions in the social media messages, the messages were read faster and rated slightly more plausible (or more likely to be acceptable in a certain situation); when the sources presented unexpected positions, the messages were read more slowly and rated slightly less plausible. Overall, the use of prior beliefs as a resource could be especially applicable when people initially browse short social media messages using quick, superficial processing (Metzger & Flanagin, 2013) before they stop to evaluate the information more strategically.

2.1.2. Discourse-Based Validation

According to Barzilai et al. (2020), evaluating arguments presented in a text is an essential component of discourse-based validation. Discursive elements include the argument's structure and coherence as well as the quality of evidence provided. Arguments may have a rhetorical structure or a dialogic structure. A rhetorical argument includes a claim (or assertion) with an accompanying justification, which differs from a dialogic argument or debate, which includes an opposition between two or more assertions (D. Kuhn, 1991). A simple rhetorical argument contains one claim with only one instance of justifying support, which is called evidence (Zarefsky, 2019) or data (Toulmin, 2003). According to Toulmin (2003), an argument can also include other components, such as a warrant, which is an often unstated but implicit inference that connects the evidence to the claim. For example, empirical research (used as evidence) entitles us to conclude that vaccines are safe (the claim) because we trust that the research has followed scientific standards

(the warrant). Social media posts with restricted lengths, such as those found on Twitter, are typically structured as simple rhetorical arguments, making them a relevant context for examining how adults attend to the coherence of a single claim, evidence, and warrant to evaluate the credibility of a simple argument.

In addition to judging the coherence of the parts of an argument, readers can also attend to the quality of evidence in an argument. There are several different types of evidence (Hornikx, 2005; D. Kuhn, 1991; Zarefsky, 2019); these types reflect how the evidence was produced and they vary in terms of their quality (C. A. Chinn et al., 2014). Four types of evidence include evidence produced from research, testimonies, anecdotes, and social consensus (Zarefsky, 2019). Evidence resulting from rigorous research processes, such as experiments, statistical analyses, and meta-analyses, can offer causal and statistical evidence. A second type of evidence is testimonial evidence (Zarefsky, 2019) or expert evidence (Hornikx, 2005). Testimony, which can consist of either a fact or an opinion, is a statement made by some source. It is reasonable for an author to rely on testimonial evidence from a qualified source to support one's claim when an author does not have direct knowledge of the topic but can trust the expertise of others.

Anecdotal evidence, such as personal experience, relies on a narrative way of knowing (Hinyard & Kreuter, 2007) while social consensus refers to widely agreed-upon facts, shared value judgments, shared historical understandings, and previously established claims (Zarefsky, 2019). Anecdotal evidence and social consensus, two types of evidence that are not necessarily supported by accurate information (e.g., research evidence or expert testimony), are common in health communication, especially in social media contexts (Suarez-Lledo & Alvarez-Galvez, 2021). Thus, social media posts are an authentic context for determining if adults can evaluate the extent to which evidence supporting certain claims provides accurate and inaccurate information.

Because social media contain an enormous number of short messages competing for readers' attention, the persuasiveness of evidence is crucial for authors to craft their arguments in ways that are both efficient and effective. When examining the relative persuasiveness of different evidence types (e.g., anecdotal versus statistical evidence), results have been contradictory. In one narrative literature review, Hornikx (2005) reviewed twelve studies that compared the persuasiveness of anecdotal evidence and statistical evidence. Six of those studies showed that statistical evidence was more persuasive than anecdotal evidence, one study showed that anecdotal evidence was more persuasive, and five studies showed no difference between the two types of evidence. Elsewhere, a meta-analysis of 15 studies concerning health communication campaigns (Zebregs et al., 2015) showed that statistical evidence had a stronger influence on beliefs and attitudes than anecdotal evidence, but that anecdotal evidence had a stronger influence on intention, which relates more closely to affective responses. Notably, a recent review found that presenting anecdotal evidence and personal experiences was one of the 12 identified persuasive techniques used in communicating online health misinformation (Peng et al., 2022).

When evaluating evidence, readers may struggle to differentiate evidence in terms of quality, especially when judging the quality of different types of research evidence. For example, List et al. (2021) found that even though undergraduate students ($N = 82$) considered anecdotal evidence considerably less convincing than other types of evidence, they struggled to differentiate between certain type of research evidence that imply causation and observational and correlational evidence (that do not imply causation). This is especially concerning because, on social media, discussions of health issues often include causal claims or generalizations that may or may not be supported by appropriate causal evidence.

2.2. Second-Hand Evaluation

The trustworthiness of a source refers to readers' perceptions of positive characteristics on the part of that source; these perceptions, allow them to accept the source's message (Fogg, 2002; Ohanian, 1990). Evaluation of source trustworthiness is essential, particularly when readers evaluate scientific information, including health information, about which they do not have specialized knowledge (Hendriks et al., 2015). Hendriks et al. (2015) found three dimensions that laypersons rely on when deciding whether they trust an expert. These dimensions are expertise, integrity, and benevolence. A source's *expertise* can be evaluated by exploring that source's credentials, such as degrees, professional achievements, affiliations, and other indicators of competence, such as awards and publications (Bråten et al., 2018; Rolin, 2020). Furthermore, *integrity* refers to honesty, objectivity, and acting according to scientific and professional standards, whereas *benevolence* refers to goodwill and intentions toward others and society (Hendriks et al., 2015).

It is widely acknowledged that information about a source's expertise plays a pivotal role in credibility evaluation of social media messages and beyond (Lin et al., 2016; Meinert & Krämer, 2022; Stadtler & Bromme, 2014). For example, Hirvonen et al. (2019) have studied cognitive authority in online health-related information-sharing processes. They

found that authority-related cues, such as user information, the authority's own experience, education, and background, were important when girls and young women evaluated the information credibility of their peers in an online forum related to health topics. However, evaluation of the source's trustworthiness can also be based on symbols of authority and fruits of success rather than on perceived expertise (Shieber, 2015). For example, in one interview study, all 18 interviewed participants (ages 18 to 30) viewed celebrities as trustworthy sources of online information (Djafarova & Rushworth, 2017).

In other research, Lin et al. (2016) studied the credibility evaluation of mock Twitter posts that discussed health risks. In the study, 696 undergraduates participated in a quasi-experiment in which the following aspects of Twitter posts were manipulated: author information, originality of the tweets (original or retweeted), and bandwagon cues (i.e., replies or not). The authors were manipulated by authority and whether the identifying information was provided. The Twitter pages were owned by either the Center for Disease Control and Preventions (CDC), a college student, or a stranger without any identifying information. After viewing the pages, students were asked to evaluate the source's credibility according to three criteria: competence, goodwill, and trustworthiness. The researchers determined that participants viewed authority cues as the most credible compared to the other cues (i.e., identity, original or retweeted, bandwagon cues). The CDC was evaluated as more credible than the student and the stranger. The bandwagon cues also had an effect, such that the student and the stranger were evaluated as more credible when they retweeted each other. Interestingly, the CDC twitter page without any retweets was perceived as the most credible in all evaluated aspects.

Some studies have examined other features of the source, such as gender and ethnicity. Spence et al. (2013) investigated intercultural differences in responding to health messages specifically in the context of Facebook. In their study, participants from the Caucasian and African American communities ($N = 200$) were evaluated to determine their response efficacy and behavioral intentions after being exposed to a social media page with an avatar. The African American page owner communicated either high or low ethnic identity (that is, including a greater or lesser number of cues on their page to promote their ethnic identity) while encouraging participants to read a story about heart disease. African American participants perceived both ethnic authors as more competent, caring, and trustworthy than Caucasian participants did. While the perceived credibility of the author was not measured, African American participants indicated stronger intention to change their dietary behavior after seeing a health message delivered by an African American avatar with high ethnic identification as compared to one with low ethnic identification.

Armstrong and McAdams (2009) examined the role of gender in credibility evaluation of blogs. They found that gender influenced how undergraduates at a large university in southeastern United States perceived credibility. The blog posts with male authors were seen as more credible than posts with female authors. Similar results regarding credibility and gender have also been attained with newscasters (Weibel et al., 2008). More recently, Groggel et al. (2019) investigated how people perceived both gender and ethnicity as well as physical attractiveness as a cue indicating trustworthiness when evaluating Twitter accounts. In the study, Amazon Mechanical Turk workers evaluated the trustworthiness of 816 Twitter profiles. Results indicated that attractiveness was positively associated with trust. Furthermore, Twitter accounts belonging to white individuals were evaluated as more trustworthy than accounts belonging to Black individuals. Also, accounts belonging to Black males and both Black and White females were viewed as less trustworthy than accounts belonging to White males. White crowdsourcing workers also evaluated accounts belonging to White individuals as more trustworthy.

2.3. Contextual Attributes of Readers

At least two studies have found that cultural context or a reader's membership in a particular social media network influenced adults' credibility perceptions of social media posts. Yang et al. (2013) compared credibility perceptions of microblog posts among U.S. and Chinese participants in two social media networks; the posts varied in terms of author's gender, name style, profile image, location, and degree of network overlap with the reader. They identified several key differences in how users from each country critically consumed microblog content in the two networks, including the observation that Chinese respondents considered information from microblogs to be more credible than U.S. respondents. However, the researchers explained "it wasn't clear whether this difference was due to inherent differences in credibility quality of microblog information in the two cultural contexts, or to the different way that people perceive the credibility of microblog information and/or other information sources." (p. 581)

In another study of 746 adults from 76 countries, S. Shariff et al. (2016) used chi-square analyses to correlate tweet credibility evaluations and demographic attributes (gender, age, education, and location) partitioned into three different settings. That is, location data were partitioned into original setting (Asia, Europe, South America, North America,

Table 1
Independent and control variables in a set of the evaluation tasks for accurate and inaccurate posts.

Variable	Description	Manipulation
Independent Variables:		
Evidence Type[1]	Evidence type supporting the main claim of the social media post.	A) Research (only for accurate claims) B) Testimony C) Personal experience D) Consensus
Prior Belief Consistency[2]	Consistency of self-rated prior beliefs about the topic in relation to the post with the accurate or inaccurate knowledge claim.	
Source[3]	The author of the social media post.	A) Professor of medicine B) Nurse education practitioner C) Professional lifestyle blogger D) Parent
Source's Gender[3]	Self-identified gender of the author presented in the profile picture.	A) Female B) Male
Source's Ethnicity[3]	Self-identified ethnicity of the author presented in the profile picture.	A) Asian B) Black C) White
Platform[4]	Crowdsourcing platform on which participants completed the set of credibility evaluation tasks.	A) Prolific B) Toloka
Control Variable:		
Topic	The topic of the health-related social media post.	A) Fish oil B) Food healthiness C) Processed red meat D) Vaccine safety E) Vitamin D

[1]First-hand evaluation: Discourse-based validation, [2]First-hand evaluation: Knowledge-based validation, [3]Second-hand evaluation, [4]Contextual attributes.

and Africa), binary setting (Eastern hemisphere, Western hemisphere) and categorical setting (Asia-Pacific, Americas, Europe, Africa). They found a combination of readers' education background and their geo-location had a significant correlation with credibility judgments, but gender and age did not. Further, only location was significantly correlated at all levels of data partitioning. Finally, S. M. Shariff (2020) conducted a literature review of credibility perceptions of online information and concluded that most studies have not paid sufficient attention to how and why readers carry out their credibility judgements of social media posts. Shariff called for more research in credibility perceptions that attends to demographic attributes and other participant characteristics (e.g., membership in particular groups or social media platforms).

Considering all of these findings, in this study, we focused on credibility judgements of health-related social media posts to determine to what extent adult readers relied on their prior beliefs (i.e., knowledge-based validation), evidence types presented in the posts (i.e., discourse-based evaluation) and source characteristics (expertise, gender, and ethnicity) after controlling for the topic of health-related posts. We also examined whether crowdsourcing platform membership played a role in readers' credibility evaluations given that the nationalities of members across the two platforms did not overlap.

3. Methodology

The present study was intended to understand how adults evaluate short social media posts on health issues. We created accurate (i.e., in line with current scientific knowledge) and inaccurate (i.e., not in line with current scientific

knowledge) social media posts on five health topics. Then, we used a computerized algorithm that systematically manipulated evidence type in the posts and source characteristics (the author's expertise, gender, and ethnicity) to generate 840 unique combinations of the posts. Each unique post was copied 10 times to generate a pool of 8400 posts. The participants, recruited from two crowdsourcing platforms, were asked to evaluate the credibility of ten different posts selected from the larger sample and delivered to each participant via the platform. We examined how evidence type, participants' prior beliefs, source characteristics of the post and crowdsourcing platform membership were associated with participants' credibility judgments on the posts after controlling for the topic of the posts.

Based on theoretical assumptions supported by previous empirical evidence (Barzilai et al., 2020; Richter & Maier, 2017; Stadtler & Bromme, 2014), we assumed that the quality of evidence, participants' prior beliefs, and expertise of the source would be associated with participants' credibility judgments. First, we expected that the more that participants' prior beliefs were consistent with the content in the post, the higher the participants would evaluate the credibility of the post (McCrudden & Barnes, 2016; Wertgen & Richter, 2023). Second, we expected that posts with claims supported by research-based evidence would be evaluated as more credible than posts that contained evidence from testimony, personal experience, or consensus (Hornikx, 2005; List et al., 2021). Third, we expected that posts by sources with relevant domain expertise would be evaluated more credible than posts by sources without domain expertise (Bråten et al., 2018; Lin et al., 2016). In addition, we were interested in testing the effect of the source's gender and ethnicity on credibility judgements as well as exploring whether membership in a certain crowdsourcing platform would be associated with participants' credibility judgments. In our analyses, we controlled for the topic of the posts because topic has been shown to play a role in the credibility evaluation of texts (Bråten et al., 2018; Hämäläinen et al., 2021).

To study these associations, we ran an online quasi-experiment with an incomplete repeated measures design (Shaughnessy et al., 2012). That is, there was no control group, and a single rater rated a maximum of 10 posts out of the 840 unique generated posts. The factors were controlled independently, with practice effects controlled with a completely random order in the Toloka platform, and with balancing in the Prolific platform. The final dataset (8380 posts) consisted of 10 ratings of the 820 unique posts from different raters, and 9 ratings for 20 of the unique posts. All the variables included in this study are presented in Table 1.

3.1. Participants

We used crowdsourcing to recruit participants for our study. The popularity of crowdsourcing has increased steadily, and it has been used in a variety of fields (Hossain & Kauranen, 2015). A total of 844 participants were recruited from two crowdsourcing platforms: Prolific and Toloka. According to Chapkovski (2023), Toloka provides access to some populations that can be hard to access through Prolific or MTurk. While the MTurk population is mostly based in the US and India (Difallah et al., 2018), and Prolific users reside mostly in the UK and US (Peer et al., 2017), Toloka's users mostly reside in Russia and other former Soviet countries. Thus, we used the Prolific and Toloka platforms to obtain a geographically and culturally broad sample.

The Prolific platform is available for participants who are at least 18 years old from most OECD countries, except for Turkey, Lithuania, Colombia and Costa Rica where Prolific is not available ((Prolific, 2023b)). Prolific is also available to participants in South Africa and other participants who are not from OECD supported countries can still participate if they live in a supported country. In September 2023, according to Prolific's audience checker tool (Prolific, 2023a), approximately 58% of its users identified as female, while 42% identified as male. Regarding age, 47% reported being under 30 years old, 48% were between the ages of 30 and 59, and 5% were 60 or older. Prolific crowdworkers are primarily recruited by word-of-mouth through social media (e.g., Facebook, Twitter, Reddit) and flyers distributed on university campuses. All participants are paid rewards when their submissions are approved.

According to its website, the Toloka platform collects data from crowdworkers who live in more than 100 different countries, with a "large concentration of participants coming from Pakistan, Kenya, Brazil, Turkey, India, Egypt, the Philippines, and the United States," and "about half of Tolokers are native speakers of English, Urdu, Arabic, Russian, or Spanish" (Toloka, 2023). The majority of Toloka workers are between 20–30 years of age; they are equally divided between members of the middle class and working class, and 62% of its members identify as male. Toloka acquires its users at least through social media presence and advertising.

Participants were required to complete a demographic survey at the beginning of the study. We used Pew Research's Demographics Questionnaire (PewResearchCenter, 2015), as the basis for our survey. As shown in Table 2, the mean age of respondents was higher in the Prolific platform ($M = 43$) compared to Toloka ($M = 33$) and there were more male respondents in Toloka (56%) than in Prolific (49%). In terms of ethnicity, a large percentage of respondents in

both platforms identified as White (73% in Toloka and 85% in Prolific), fewer identified as Asian (11% in Toloka and 5% in Prolific) or as individuals representing multiple ethnicities (10% in Toloka and 7% in Prolific) while the smallest number of respondents in either platform identified as Black (6% in Toloka and 2% in Prolific). Education levels were similar across the two platforms, but notably, there was no apparent overlap in the nationalities of respondents across the two platforms. In Toloka, more than half (56%) of respondents identified as Russian, while the remaining respondents lived in regions in Eastern Europe, the Middle East, Kenya, or India. In contrast, more than three-quarters (77%) of respondents in Prolific were from either the United Kingdom (65%) or the United States (12%), and the remaining respondents were from either Canada or other countries in Western and Southern Europe.

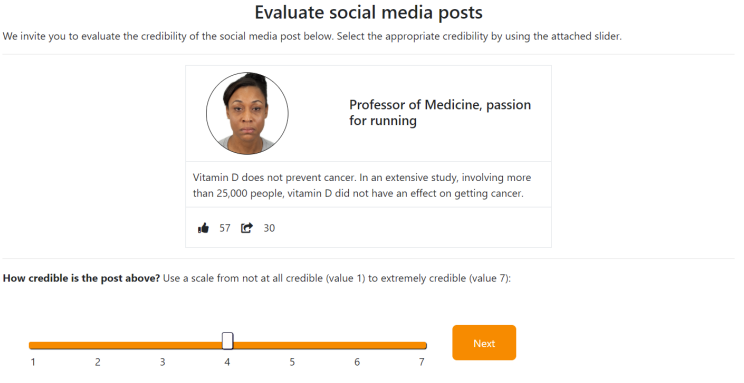
3.2. Generation of Credibility Evaluation Items for Social Media Posts

3.2.1. Designing Posts

We designed short social media posts regarding five health topics on which contradictory information, including accurate and inaccurate claims, had been spread on the internet. In designing the social media posts, we consulted a current Finnish book written by Knuuti (2020), a professor and director of the Turku PET Centre, University of Turku. For this book, Knuuti collected 250 common health-related knowledge claims spreading on the internet and analyzed them in light of current scientific knowledge. The chosen topics were the safety of vaccines, Vitamin D, processed red meat, fish oil, and food healthiness. There are several false claims about vaccines on the internet, one of which is that the aluminum in vaccines is dangerous. Also, excessive claims about the benefits of Vitamin D, such as preventing cancer, have been presented. Furthermore, research on the health effects of red processed meat has been misinterpreted by neglecting the potential harms. Finally, there are contradictory claims about the benefits of fish oil and food healthiness. We used research presented in the book to design the research-based social media posts. The sources provided in the book are presented in Table 7 of Appendix.

Each post comprised three main components which were systematically manipulated with a computer algorithm: the knowledge claim, supporting evidence, and author of the post (see Figure 1). Knowledge claims were either accurate, that is, in correspondence with the current scientific knowledge (e.g., Vitamin D does not prevent cancer), or inaccurate (e.g., Vitamin D prevents cancer). The knowledge claims were paired with supporting evidence that represented one of the following evidence types: research evidence, testimony, personal experience, and social consensus (Jacobsen et al., 2018; Zarefsky, 2019). Because we used authentic findings as research evidence, we only presented these findings with accurate claims. Examples of the knowledge claims and evidence types related to processed red meat are presented in Table 3.

Figure 1: Example of the credibility evaluation task item on the Prolific platform.



We created four types of authors (or sources) and varied their expertise. Two sources, a professor of medicine and a nurse education practitioner for undergraduate students, were experts in the health domain. The other two sources, a professional lifestyle blogger and a parent, did not have domain expertise. In addition to this information, the author's profile included a short description, such as 'passion for running', to make the social media messages seem more authentic.

Because some prior work has suggested that an author's ethnic background and gender may be associated with readers' responses to social media health information (Armstrong & McAdams, 2009; Spence et al., 2013), we also

Table 2
Participants self-reported demographic information by crowdsourcing platform.

	Toloka	Prolific	Total
Respondents	425	419	844
Mean age	33	43	38
Minimum age	18	20	
Maximum age	85	84	
25 or under	107 (25%)	25 (6%)	132 (16%)
26-35	139 (33%)	127 (30%)	266 (32%)
36-45	108 (25%)	100 (24%)	208 (25%)
46-55	36 (8%)	83 (20%)	119 (14%)
over 55	35 (8%)	84 (20%)	119 (14%)
Gender:			
Male	240 (56%)	204 (49%)	444 (53%)
Female	180 (42%)	215 (51%)	395 (47%)
Other	5 (1%)	0	5 (<1%)
Ethnicity:			
Asian	47 (11%)	23 (5%)	70 (8%)
Black	26 (6%)	10 (2%)	36 (4%)
White	309 (73%)	357 (85%)	666 (79%)
Other or multiple	43 (10%)	29 (7%)	72 (9%)
Nationalities:	Russia – 247 (56%)	UK – 272 (65%)	UK – 273 (32%)
	Ukraine – 27 (7%)	United States – 49 (12%)	Russia – 247 (29%)
	Turkey – 21 (5%)	Italy – 15 (4%)	United States – 56 (7%)
	Belarus - 20 (5%)	Spain - 11 (3%)	Ukraine - 27 (3%)
	Pakistan - 15 (4%)	Poland - 11 (3%)	Turkey - 23 (3%)
	India - 15 (4%)	Canada - 11 (3%)	Belarus - 20 (2%)
	Kenya - 13 (3%)	Portugal - 9 (2%)	India - 15 (2%)
	Other - 67 (16%)	Other - 41 (10%)	Other - 183 (22%)
Education:			
Doctorate	5 (1%)	11 (3%)	16 (1%)
Master's degree	73 (17%)	68 (16%)	141 (17%)
Bachelor's degree	148 (35%)	173 (41%)	321 (38%)
High school diploma	107 (25%)	123 (29%)	230 (27%)
Professional degree	59 (14%)	17 (4%)	76 (9%)
Other qualifications or none	33 (8%)	27 (6%)	60 (7%)

varied the author gender and ethnicity. To do this, we used the Chicago face database (Ma et al., 2015), which provides facial pictures along with the self-identified ethnicity and gender, as well as a subjective attractiveness score based on ratings from US-based raters. Using this data, we selected pictures of people with average attractiveness who were between 35 and 45 years of age. We included males and females, as well as those who self-identified as Asian, Black, and White. The age and attractiveness selection were performed to create a uniform set of facial pictures and thus rule out very unlikely combinations, such as 20 year old professors. The selected profile pictures are shown in Appendix (Figure 5).

3.2.2. Full Factorial Design

Next, unique combinations of sourcing features (author expertise, gender, ethnicity, and profile picture) and content features (topic, claim type, and evidence type) were generated by applying mathematical principles of combinatorics (R. Kuhn et al., 2009) to a full factorial design to study all possible combinations of the different levels of factors being considered. Traditional experimental designs in credibility evaluation often avoid full factorial design because the total number of all combinations can escalate quickly, surpassing what is feasible to study - hence, the complexities we touch upon in the paper title. For instance, an experiment may examine two factors: source expertise of a post and evidence backing the post. If each factor has two levels (e.g., source expertise includes 1) Professor and 2) Blogger, and evidence type comprises 1) Research and 2) Personal experience), this results in four unique combinations (2x2). If

Table 3
Arguments of the evaluated social media posts concerning processed red meat.

Argument with the accurate claim	Evidence type
Red processed meat is a major health hazard. The study review found that red processed meat was associated with a higher incidence and mortality of arterial disease.	Research
Red processed meat is a major health hazard. The WHO places processed meat into the highest risk category of carcinogens..	Testimony
Red processed meat is a major health hazard. Humans with a balanced diet without red processed meat feel better.	Consensus
Red processed meat is a major health hazard. I eat red meat regularly and at my physical, some of my important indicators are alarming.	Personal Experience
Argument with the inaccurate claim	
Red processed meat is not a significant health hazard. My personal trainer appreciates red processed meat as a part of a balanced diet.	Testimony
Red processed meat is not a significant health hazard. Humans are carnivores and this is how we have survived.	Consensus
Red processed meat is not a significant health hazard. I eat red meat regularly and at my physical, all of my important indicators are great.	Personal Experience

there are three levels for each factor, then nine (3x3) combinations arise. Introducing a third factor with three potential values, such as source ethnicity, balloons the number of combinations to 27 (3x3x3). The number of factors and levels renders full factorial design seemingly unfeasible in studies of credibility evaluation.

In our study, we implemented a full factorial design that tested 840 unique combinations. With such a vast number of combinations, we encountered two main challenges. The first was the effort of creating 840 distinct social media posts and ensuring that each combination was created just once. The second challenge was how to conduct an experiment with this multitude of posts. We addressed the first issue using an algorithm and software that automatically generated the required 840 social media posts, ensuring all combinations were produced. This resulted in 480 unique posts with accurate knowledge claims and 360 unique posts with inaccurate knowledge claims. For the second challenge, we employed two crowdsourcing platforms, from which large numbers of participants could easily be recruited. To conduct the experiment, we used an incomplete repeated measures design since expecting one person to evaluate all 840 posts would be unrealistic. As a result, each participant only assessed 10 unique posts.

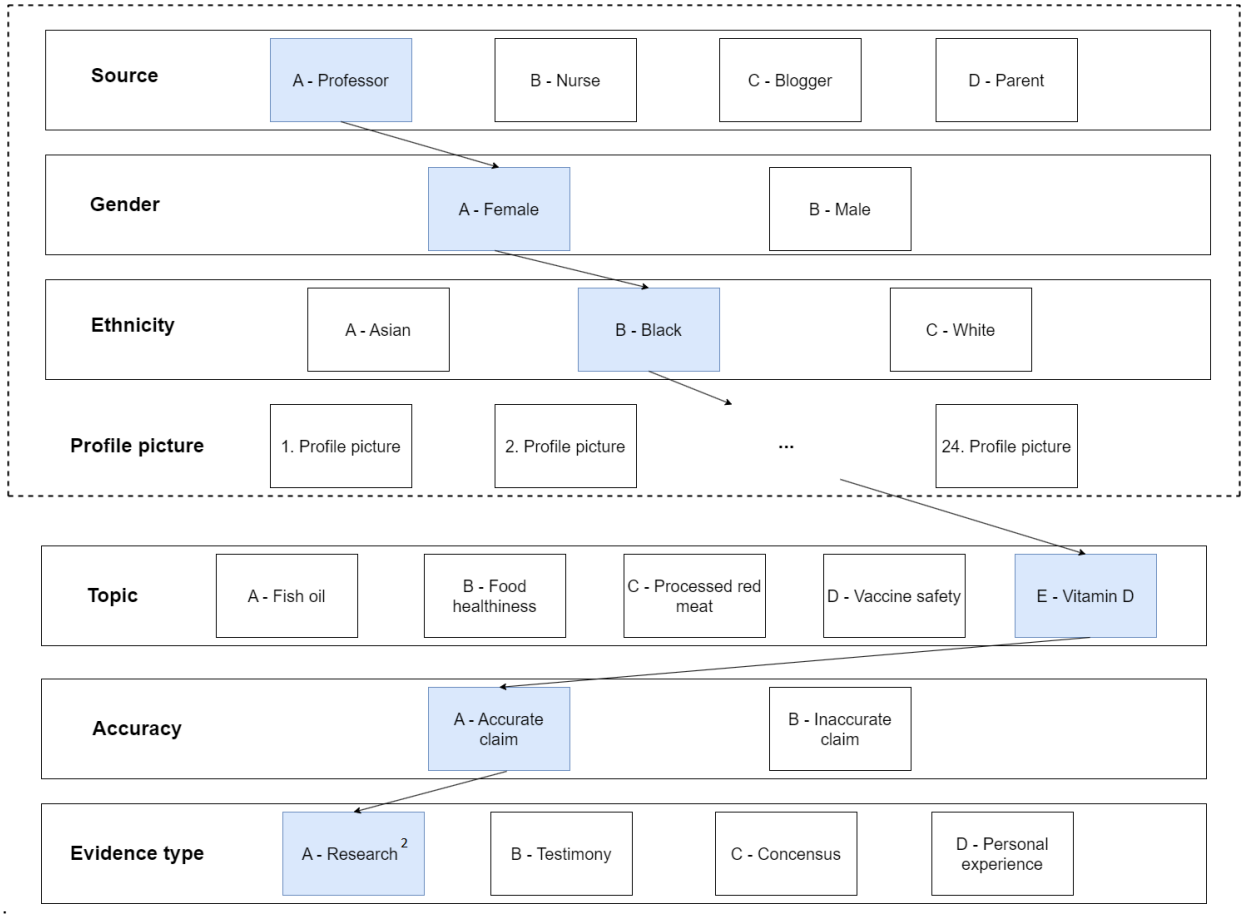
Figure 2 illustrates how social media posts were generated with this computerized procedure by using the credibility evaluation task item presented in Figure 1 as an example. A combination of ethnicity, gender, and profession was used to select the profile picture (i.e., Black female professor) associated with the social media post, which was generated using a combination of the topic (Vitamin D), accuracy of claim (accurate), and evidence type (research).

Although we conceptualized perceived credibility as a multidimensional construct, reflecting the idea of first- and second hand evaluation (Barzilai et al., 2020; Stadtler & Bromme, 2014), we measured participants' credibility judgments with a single item (cf. Hanimann et al., 2023). Participants were assigned to evaluate the credibility of ten posts on a single item scale from 1 to 7 (1 = not at all credible; 7 = extremely credible). The single item was used because we sought to understand how different content- and source-related issues affected participants' overall evaluations of the social media posts' credibility. Single-item scale may be advantageous because participants may resent being asked questions that appear to be repetitive (Fuchs & Diamantopoulos, 2009). In this study, the unit of analysis was the credibility evaluation of one post.

3.3. Prior Belief Measure

Before gaining access to the credibility evaluation task, participants' prior beliefs on the topic were measured using one statement for each topic at hand and a 7-point scale. The statements were worded, for example, 'Vaccines are safe', and 'Red meat is healthy'. We coded the responses to prior belief items in relation to how consistent they were with the

Figure 2: Illustration of algorithmic process[1] used to generate 840 unique combinations of social media posts.



:

Note: [1] Rows 1-3 represent steps in the computerized process of selecting a unique combination of source features (source, gender, and ethnicity [i.e., Professor who is female and Black] paired with a profile picture matching those features. Rows 4-6 represent steps in generating content of the social media post associated with that source by selecting a topic (Vitamin D), claim type (accurate), and evidence type (research).

[2] The Research type of evidence was only available for accurate claims.

claim. For accurate claims, a prior belief score of 7 corresponded to a prior belief consistency score of 3, and a prior belief score of 1 corresponded to a consistency score of -3. For inaccurate claims, this was done in reverse, with a prior belief score of 1 corresponding to consistency score of 3 and a prior belief score of 7 corresponding to a consistency score of -3. For example, if the participant responded with 1 on the item "Vaccines are safe,", it was coded as -3 for the accurate claim. In the case of an inaccurate claim, the participant's response 1 was coded as 3. The variable was labeled as prior belief consistency, with a scale from -3 to 3.

3.4. Data Collection Procedure

Prior work has established cultural differences in credibility evaluation of online reviews (Brandt, 2022), as well as how people use and trust health information in online contexts (Khosrowjerdi, 2020; Song et al., 2016). Because we opted for two crowdsourcing platforms, with a wide variety of nationalities in the userbase, crowdworkers on Toloka and Prolific were offered an opportunity to take part in our study. Participants were first asked to fill out a questionnaire on demographic information and the prior belief measure. After completing the questionnaire successfully, the participants proceeded to the credibility evaluation task, in which they were asked to evaluate the credibility of ten social media posts. We assigned only ten credibility evaluation items to each participant to ensure

that they could focus sufficiently on each item. In addition, with this procedure, we could recruit a considerable number of participants which assured that they represented diverse backgrounds.

On the two platforms, somewhat different methodologies were used to deliver the set of credibility evaluation items to raters because of how these platforms worked. On the Prolific platform, we gave the tasks to participants in predetermined batches of ten, with minimized repetition of the manipulated aspects. In principle, ten posts could have some repetitive elements. The same topic, evidence type, source, and profile picture could have been shown to the participant in a maximum of three posts. However, the exact same post was not shown twice to the same participant. On the Toloka platform, each participant had an opportunity to make ten evaluations that were assigned to them at random. This was done by randomizing the order of posts that were given to the crowdworkers on a first-come first-serve basis. However, like Prolific, the exact same post was not shown twice to the same participant. Thus, the items were manipulated independently of each other. Only the top 20 % performing crowdworkers on the Toloka platform were offered the chance to participate in the study. Crowdworkers' performance was calculated by their success in completing tasks on the platform (i.e., the acceptance rate of their previous work at the platform). All evaluations made on the Toloka platform were accepted. Of the Toloka participants, 99% completed the maximum of ten evaluations. The task used on the Prolific platform is shown in Figure 1.

3.5. Statistical Analysis

In the statistical analysis, the unit of analysis was the credibility evaluation of one social media post. Analyses were conducted separately for posts including accurate and inaccurate knowledge claims, which are later referred to as accurate and inaccurate posts below. Data consisted of 4787 credibility evaluations of accurate posts and 3593 evaluations of inaccurate posts. Twenty credibility evaluations were excluded from the analysis due to technical problems.

All statistical analyses were conducted using R. To examine how independent variables (see Table 1) were associated with the credibility evaluations of social media posts, we used a cumulative link mixed model. The cumulative link mixed model (CLMM), also known as ordered regression model (Christensen, 2015), has been advocated for use with ordered outcome measures (Fullerton & Anderson, 2023).

To provide a CLMM model and odds ratios, we used the ordinal package created by Christensen (2015) for the R programming language. The specific formula used is follows: "Credibility~Accuracy + ClaimEvidence + Topic + Author + Platform + Gender + Ethnicity + (1|ID)". The control variable ID corresponds to each individual evaluator who performed a maximum of ten evaluations in the sample. For the reference category, that is, the category other categories are compared to in the model (see reference categories in Tables 5 and 6), we chose the category that one would assume to produce the highest credibility rating. For all other variables, we used the category that was first in alphabetical order. Thus, for evidence type, we chose research; for prior beliefs, we chose highest consistency, and for source expertise, we chose professor. For gender, ethnicity, platform, and topic, we used alphabetic order.

We also present odds ratios (ORs), which are calculated with the ordinal package (Christensen, 2015), to highlight the association between the independent variable and the dependent variable. Odds ratios have been advocated for use as an effect size measure with cumulative link mixed models and ordered regression models (Fleiss, 1994; Fullerton & Xu, 2016). Moreover, odds ratios have been used with CLMM models in a variety of fields (Oehm et al., 2022; Rezapour & Ksaibati, 2021). An OR greater than 1 indicates how many times greater the expected credibility evaluation value of the category of interest is compared to the reference category. In contrast, an OR smaller than 1 indicates that the expected evaluation value in the reference category is greater than that in the category of interest. The ORs were transformed to another popular effect size measure, Cohen's d, with the effectsize R package created by Ben-Shachar et al. (2020). The Cohen's d effect sizes were interpreted as follows: small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$)(Cohen, 2013). To calculate the goodness-of-fit statistics, we used the anova.clmm function of the RVAideMemoire package (Hervé & Hervé, 2020) to perform a likelihood-ratio test (Crainiceanu & Ruppert, 2004).

4. Results

In this study, we were interested in how content-related issues (i.e., prior belief consistency and evidence type), source characteristics (i.e., source expertise, ethnicity and gender), and crowdsourcing platform (i.e. Toloka or Prolific) were associated with reader's credibility evaluations of short social media posts when the posts' topic were controlled for. Specifically, we were interested in the relative impact of the identified factors on the perceived credibility of social media posts containing accurate or inaccurate health claims. Descriptive statistics are shown in Table 4. We ran the

Table 4
Descriptives for the credibility evaluations by independent and control variables.

Variable	Manipulation	Mean[1]	SD	N (8380)[2]
Accuracy of the Claim	Accurate	4.25	1.72	4787
	Inaccurate	3.65	1.81	3593
Evidence Type[3]	Research	4.36	1.64	1193
	Testimony	4.13	1.77	2400
	Consensus	3.94	1.81	2389
	Personal experience	3.74	1.80	2398
Prior Belief Consistency[4]	3	4.30	1.90	1120
	2	4.18	1.71	1162
	1	4.16	1.65	1276
	0	4.02	1.69	1583
	-1	3.82	1.73	1198
	-2	3.68	1.78	1083
	-3	3.75	2.02	958
Source[5]	Professor of medicine	4.40	1.77	2094
	Nurse education practitioner	4.22	1.69	2095
	Professional lifestyle blogger	3.80	1.78	2093
	Parent	3.56	1.78	2098
Source's Gender[5]	Female	4.02	1.80	4191
	Male	3.97	1.77	4189
Source's Ethnicity[5]	Asian	3.97	1.79	2797
	Black	4.05	1.77	2789
	White	3.97	1.79	2794
Platform[6]	Prolific	3.68	1.75	4190
	Toloka	4.32	1.77	4190
Topic	Food healthiness	4.22	1.76	1673
	Vitamin D	4.04	1.84	1680
	Fish oil	4.01	1.68	1676
	Processed red meat	3.99	1.74	1678
	Vaccine safety	3.72	1.87	1673

Note. [1]Credibility evaluations were given on a scale from 1 to 7. [2]N is the number of credibility evaluations within each variable.
[3]First-hand evaluation: Discourse-based validation, [4]First-hand evaluation: Knowledge-based validation, [5]Second-hand evaluation, [6]Contextual attributes.

CLMM analyses separately for credibility evaluations of the posts with accurate and inaccurate knowledge claims. A Likelihood-ratio test between full models and their null model counterparts, that is, models without predictors, showed a good fit, with both being statistically significant at the $p < .001$ level.

Results of the CLMM analysis are presented in Table 5 for the posts with accurate claims and in Table 6 for the posts with inaccurate claims. As assumed, the content-related issues (i.e., prior belief consistency and evidence type) affected readers' credibility evaluations of the posts regardless of whether the claims were accurate or inaccurate. As shown in Tables 5 and 6, participants' prior beliefs significantly affected their credibility evaluations. The more consistent participants' prior beliefs were with the claim of the post, the more the message was evaluated as being credible. Effect sizes (Cohen's d) between evaluations of the very high consistency posts and the very low consistency posts were -0.71 for posts with accurate claims and -1.17 for posts with inaccurate claims.

The role of evidence type in credibility evaluation was small. When other evidence types were compared to research evidence (posts with accurate claims), the effect sizes ranged from 0.11 to -0.18, and when compared to testimony (posts with inaccurate claims), the effect sizes ranged from 0.01 to -0.18. Notably, for the posts with accurate claims, the posts with testimony were evaluated the most credible, testimony being 1.21 times more likely to be evaluated credible

Table 5
CLMM results for accurate social media posts.

Fixed effect terms	Estimate	Std. error	Odds Ratio	Cohen's d
Independent variables:				
Evidence Type[1] – Reference: Research				
Testimony	0.19*	0.08	1.21	0.11
Consensus	-0.29***	0.08	0.75	-0.16
Personal experience	-0.33***	0.10	0.72	-0.18
Prior Belief Consistency[2] – Reference: 3				
2	-0.45***	0.12	0.64	-0.25
1	-0.68***	0.12	0.50	-0.37
0	-1.19***	0.11	0.30	-0.65
-1	-1.21***	0.13	0.30	-0.66
-2	-1.48***	0.15	0.23	-0.82
-3	-1.28***	0.14	0.28	-0.71
Source's Expertise[3] – Reference: Professor of medicine				
Nurse education practitioner	-0.28***	0.08	0.75	-0.16
Professional lifestyle blogger	-0.89***	0.08	0.41	-0.49
Parent	-1.13***	0.08	0.32	-0.62
Source's Gender[3] – Reference: Female				
Male	-0.04	0.06	0.96	-0.02
Source's Ethnicity[3] – Reference: Asian				
Black	0.10	0.07	1.11	0.06
White	0.08	0.07	1.08	0.04
Platform[4] - Reference: Prolific				
Toloka	0.69***	0.09	2.00	0.38
Control variable:				
Topic – Reference: Fish oil				
Food healthiness	1.49***	0.09	4.44	0.82
Processed red meat	0.69***	0.09	1.99	0.38
Vaccine safety	0.39***	0.09	1.47	0.21
Vitamin D	1.57***	0.11	4.80	0.86
Post-Hoc analysis – Evidence Type:				
Concensus – Testimony	-0.48***	0.08	0.62	-0.26
Concensus – Personal experience	0.04	0.08	1.04	0.02
Personal experience – Testimony	-0.52***	0.09	0.60	-0.28
Post-Hoc analysis – Source's expertise:				
Nurse education practitioner – Professional lifestyle blogger	0.60***	0.08	1.82	0.33
Nurse education practitioner – Parent	0.85***	0.08	2.34	0.47
Parent – Professional lifestyle blogger	-0.25***	0.08	0.80	-0.14

Note. ** $p < 0.01$ ***$p < 0.001$. [1]First-hand evaluation: Discourse-based validation, [2]First-hand evaluation: Knowledge-based validation, [3]Second-hand evaluation, [4]Contextual attributes.

than research evidence. When evaluating posts with inaccurate claims, readers evaluated the posts with testimony and consensus similar in credibility.

In terms of the source characteristics, the expertise of the source affected credibility evaluations, whereas source gender and ethnicity did not have any effect. The posts written by the expert sources, that is professor of medicine or nurse education practitioner, were evaluated to be more credible than posts by non-expert sources. Cohen's d varied from -0.43 to -0.62 when the expert sources were compared to the parent and from -0.39 to -0.49 when the expert sources were compared to the professional lifestyle blogger.

In terms of contextual attributes of the crowdworkers themselves, the choice of data collection platform had a notable impact on our results. Participants from the Toloka platform tended to perceive both kinds of posts as more

Table 6
CLMM results for inaccurate social media posts.

Fixed effect terms	Estimate	std. error	Odds Ratio	Cohen's d
Independent variables:				
Evidence Type[1] – Reference: Testimony				
Consensus	0.01	0.09	1.01	0.01
Personal experience	-0.32***	0.10	0.72	-0.18
Prior Belief Consistency[2] – Reference: 3				
2	-0.23	0.17	0.79	-0.12
1	-0.61***	0.17	0.54	-0.34
0	-0.85***	0.16	0.43	-0.47
-1	-1.31***	0.17	0.27	-0.72
-2	-1.54***	0.17	0.21	-0.85
-3	-2.12***	0.18	0.12	-1.17
Source's Expertise[3] – Reference: Professor of medicine				
Nurse education practitioner	-0.08	0.09	0.93	-0.04
Professional lifestyle blogger	-0.71***	0.10	0.49	-0.39
Parent	-0.86***	0.10	0.42	-0.47
Source's Gender[3] – Reference: Female				
Male	-0.10	0.07	0.91	-0.05
Source's Ethnicity[3] – Reference: Asian				
Black	0.11	0.08	1.12	0.06
White	-0.07	0.08	0.93	-0.04
Platform[4] – Reference: Prolific				
Toloka	1.04***	0.11	2.83	0.57
Control variable:				
Topic – Reference: Fish oil				
Food healthiness	-1.32	0.11	0.27	-0.73
Processed red meat	-0.90	0.11	0.41	-0.50
Vaccine safety	-1.46	0.11	0.23	-0.80
Vitamin D	-2.21	0.16	0.11	-1.22
Post-Hoc analysis – Evidence Type:				
Consensus – Observation	0.35***	0.08	1.42	0.19
Post-Hoc analysis – Source's expertise:				
Nurse education practitioner – Professional lifestyle blogger	0.61***	0.09	1.85	0.34
Nurse education practitioner – Parent	0.81***	0.09	2.24	0.44
Parent – Professional lifestyle blogger	-0.19	0.09	0.82	-0.11

Note. ***p $<$ 0.001. [1]First-hand evaluation: Discourse-based validation, [2]First-hand evaluation: Knowledge-based validation, [3]Second-hand evaluation, [4]Contextual attributes.

credible, whether they contained accurate or inaccurate claims. This disparity between the platforms was even more pronounced for posts containing inaccurate claims (d = 0.57), with a post of this nature being 2.83 times more likely to be rated higher on the Toloka platform compared to Prolific platform.

To sum up, our results suggest that adults' prior beliefs play a considerable role in the credibility evaluation of short social media posts across different sources (See Figure 3). This is particularly significant when adults face inaccurate information on social media. Furthermore, adult readers seem to pay more attention to source features indicating the source's expertise than to the quality of the evidence (See Figure 4). For example, regardless of evidence type used in the arguments, participants trusted the posts by the professor of medicine more than those by the parent. Finally, adults' credibility evaluations differed significantly across the two crowdsourcing platforms, such that crowdworkers in the Toloka platform tended to evaluate the credibility of the inaccurate posts, in particular, higher than crowdworkers in the Prolific platform.

Figure 3: Credibility evaluations for accurate (left) and inaccurate (right) posts by source and prior belief consistency.

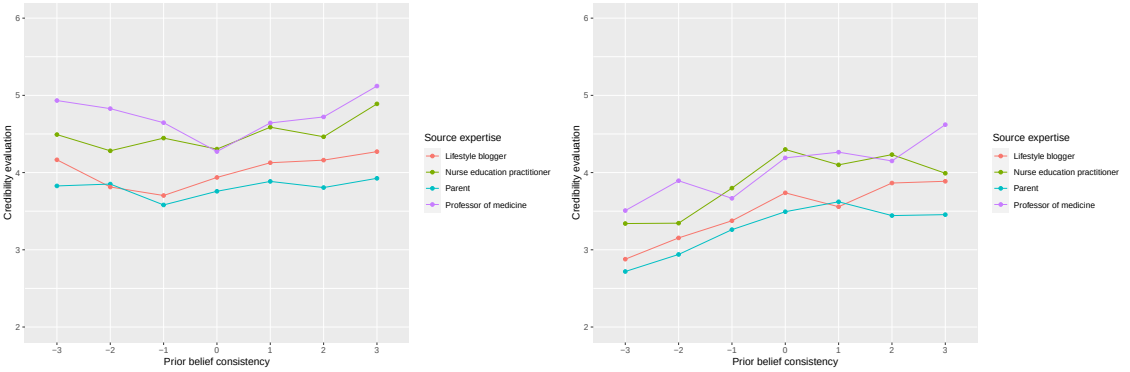
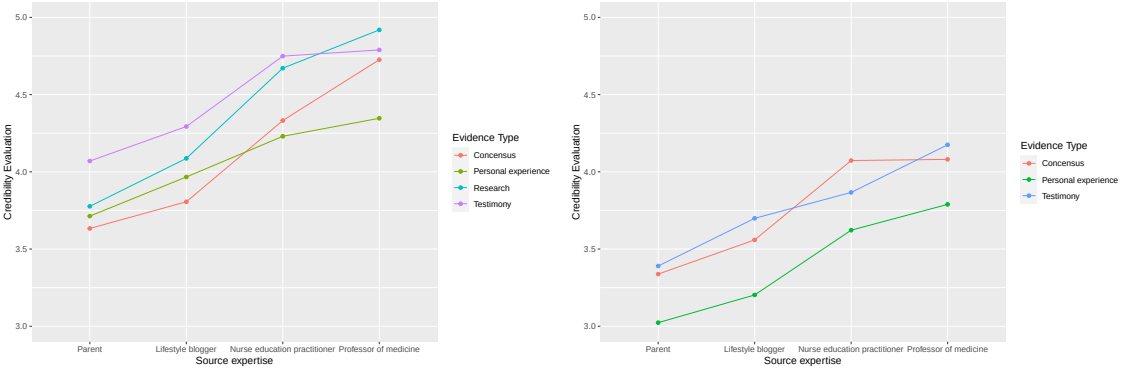


Figure 4: Credibility evaluations for accurate (left) and inaccurate (right) posts by source and evidence type.



5. Discussion

This study sought to understand how adult readers evaluate the credibility of short health-related social media posts that vary in source characteristics and argument quality. Our design of the social media posts included some unique features. First, we chose topics on which contradictory claims, including accurate and inaccurate information, spread on the internet, by relying on the thorough investigation of a medical professor who has collected scientific evidence to combat the common misinformation on health issues (Knuuti, 2020). Second, this study was the first to manipulate evidence type (i.e., research evidence, testimony, consensus, and personal experience) in social media posts showing that the quality of evidence had a minor effect on adult readers' credibility evaluations. Another unique feature of this study was that we recruited participants from two different crowdsourcing platforms to have a culturally broad sample, which had interesting results. First, we discuss the findings in terms of first-and second-hand evaluation (i.e., content and source evaluation), after which we discuss effects of the crowdsourcing platform as a context for evaluating social media posts. Finally, we discuss how these findings provide important contributions to theoretical models and methodologies for studying credibility evaluation and consider limitations of the study.

In terms of first-hand evaluation strategies, participants' prior beliefs mattered more regarding their credibility judgments than the quality of the evidence presented in the post. The more consistent the participants' prior beliefs were with the claim of the post, the more credible the post was evaluated as being. This association was found in credibility judgments of both accurate and inaccurate posts, and it is in line with previous research (Abendroth & Richter, 2021; McCrudden & Barnes, 2016). Relying on one's prior beliefs in credibility evaluation is, however, a double-edged sword. When a person has accurate prior beliefs, they can be effectively used to disregard misinformation. In contrast, when a person has false prior beliefs, relying on them in credibility evaluation is unwise, as this strategy may do nothing but strengthen one's false beliefs. Therefore, it is essential that, in our increasingly complex knowledge society, laypeople acknowledge the limits of their prior knowledge and beliefs when consuming online information (Osborne & Pimentel,

2022). This becomes even more important when comprehending health information in social media contexts, given the level of domain knowledge required to understand complicated medical terminology and evidence that misinformation on major public health issues is very common on social media platforms (Suarez-Lledo and Alvarez-Galvez, 2021; see also Vosoughi et al., 2018).

Readers' prior beliefs are a quick internal resource that they can easily access and use in their credibility evaluations. In contrast, using discourse-based validation strategies, such as evaluating the strength of the argument, may require more processing effort. However, it is still somewhat surprising that the evidence type had relatively small effect (*d*s 0.02–0.28) on participants' evaluation of accurate social media posts in which all four evidence types (i.e., research, testimony, consensus, and personal experience) were represented. It is understandable that research and testimony had a similar effect on credibility evaluation because the testimonies used in our posts relied on authorities (e.g., European Food Safety) or experts (e.g., a doctor), but it is surprising that the posts relying on consensus or personal experience were valued rather similarly to the posts that included research evidence or expert testimony.

It is true that findings in previous studies by List et al. (2021) and Hornikx (2005) indicate that research-based evidence is highly valued among readers. However, in these studies participants evaluated evidence in more traditional contexts, such as newspaper stories (List et al., 2021), not social media. Our study uniquely examined evidence as part of social media posts, suggesting that research-based evidence may be less valued in the social media context. One reason may be that narrative ways of knowing, such as personal experience, are more persuasive in social media. Instead of seeking research-based evidence for decision-making, people may turn to other consumers' or influencers' experiences that they can relate to (Feng et al., 2021). Such behavior may also be reflected in the credibility evaluation of the social media posts. Another explanation is that many students leave high school unable to understand, evaluate, or write arguments, and especially scientific arguments (Duschl & Osborne, 2002; Larson et al., 2009; Takao & Kelly, 2003) Strategies for comprehending and evaluating arguments are not explicitly taught in school and are rarely emphasized in university curricula (Osborne, 2010) and deficits in aspects of scientific reasoning are pervasive among university students (Britt et al., 2014). Thus, more research is needed to understand if increased exposure to arguments, and explicit instruction and practice in the skills of argumentation, influence adults' discourse-based credibility evaluation skills, particularly in social media contexts.

In terms of second-hand evaluation strategies, we manipulated the source's expertise, gender, and ethnicity. Our results suggest that participants focused primarily on a source's expertise when evaluating the credibility of social media posts. In fact, a source's gender and ethnicity were not associated at all with participants' credibility judgments. Our findings about the role of source expertise in credibility evaluation are in line with previous research (e.g., Lin et al. (2016)). Participants evaluated the expert sources (i.e., the professor of medicine and nurse education practitioner) as being more credible than the laypersons (i.e., the lifestyle blogger and parent) almost regardless of the presented evidence (Figure 4). This could be explained based on default trust: the default trust occurs when there is no reason to doubt the argument of the source. This kind of trust is typically based on inductive reasoning (Shieber, 2015). For example, trust in the professor's argument is based on the experience that professors typically provide accurate information. Trust in epistemic authority can also be based on monitoring for the trappings of competence, such as professional titles. This kind of monitoring is more about the symbols of authority than substance (Shieber, 2015).

In our study, a source's gender and ethnicity did not predict participants' credibility judgments. This is in contradiction to some previous work (Armstrong & McAdams, 2009; Groggel et al., 2019). In these previous studies, the focus was more on the source's gender and/or ethnicity as such. For example, when examining the gender effects on perceived credibility of blog authors, the content of the blog was kept constant (Armstrong & McAdams, 2009). In our study, the expertise of the source and content of the post were varied, which might explain why gender and ethnicity of the source did not predict participants' credibility judgments. Thus, it is possible that this variation overrode the effect of the source's gender and ethnicity.

It is worth noting that the credibility evaluation of social media posts is cognitively demanding as readers must keep several content and source features in their minds when processing and judging the posts. In this light, our results align with the limited capacity model of mediated message processing (Lang, 2000). The model suggests that due to limited cognitive capacity, people cannot process all aspects of messages they receive, and thus, readers select and process only some features of messages. Furthermore, building on this assumption, the dual processing model of credibility assessment emphasizes that readers' motivation and ability determine whether and to what degree readers will evaluate the credibility of online information (Metzger & Flanagin, 2015). Our credibility evaluation task did not have high personal relevance for the participants, crowdworkers. Thus, crowdworkers' motivation to deeply consider

a broader range of source and content features may have been low, leading to using cognitively and temporally cost-effective evaluation strategies. For example, crowdworkers may have based their credibility evaluations mainly on their prior beliefs and source expertise, neglecting the evaluation of the quality of evidence. We assume that readers' behavior would likely be somewhat different if they were engaged in seeking information needed to make important health-related decisions (cf. Sperber et al., 2010). Thus, the results of this study should be interpreted carefully.

Regarding contextual attributes of the crowdworkers themselves, because there appeared to be no overlap between the self-reported nationalities of study participants in the Prolific and Toloka platforms, we explored the possibility that cultural background (i.e., groups of individuals with nationalities associated with each platform) played a role in predicting the quality of adults' credibility evaluation practices. Notably, we found that the crowdsourcing platform in which participants completed the credibility evaluation tasks had a significant effect on their performance. Toloka crowdworkers, representing nationalities including Russia, Eastern Europe, the Middle East, Keyna, and India, tended to evaluate the credibility of the inaccurate posts, in particular, higher than Prolific crowdworkers, most of whom lived in the United Kingdom, United States, Canada, and Western and Southern Europe. One reason behind these results might be that adults living in open Westernized countries, like the participants in Prolific, may be exposed to more opportunities to critically question the quality of information sources while participants in Toloka were living in countries like Russia and Turkey with governments that are more likely to repress public forms of criticism and adopt greater restrictions targeting social media use (cf.Tufekci, 2014).Thus, Toloka participants may have less practice in negotiating alternative viewpoints and distinguishing between accurate and inaccurate information in social media posts.

5.1. Implications for Credibility Evaluation Theory and Research

Findings from this study have implications for theory and research involving the study of credibility evaluation. First, our results affirm the bi-directional model of first- and second-hand evaluation strategies (Barzilai et al., 2020) by showing that both content and source related issues were associated with reader's credibility judgments of social media posts. In addition, by accounting for multiple factors (i.e., variations in content and evidence type, variations in the source's expertise, gender and ethnicity, and differences in crowdsourcing platform membership) at the same time, rather than assigning participants to predefined conditions involving fewer factors, this study provides a richer and more complex understanding of the order of importance each factor can play in adults' overall credibility evaluations of social media posts. For example, while we learned that adults' rely on their prior beliefs and source expertise more than other source features, such as gender and ethnicity, to evaluate overall credibility, we found that evidence quality doesn't matter as much as we would hope when adult readers consume information about controversial health issues on social media. Our study also provides more conceptual clarity in how variations in these factors can be operationalized and incorporated into the design of authentic, critical online reading tasks for adults.

These findings also complexify bi-directional theories of credibility evaluation (Barzilai et al., 2020; Richter & Maier, 2017; Stadtler & Bromme, 2014) by hinting at how readers may attend to contextual attributes of the text as a third tier of credibility evaluation (Forzani et al., 2022). That is, as readers are expected to make inferences about the social and political contexts of health-related social media posts (e.g., unvetted information often paired with testimonials and endorsements by strangers), it is possible that their reasoning may need to extend beyond the use of explicit content and source features to evaluate the quality of unstated warrants in these contexts. Of course, the design of our study does not allow us to explore this possibility. To understand how participants employ first- or second-hand evaluation strategies or if there is a need for additional tiers of strategy use, methods (e.g., think-aloud, open-ended questions) allowing participants to explain their reasoning behind their credibility judgments are needed. Furthermore, more controlled experiments could clarify the nature of the reciprocal relationship between content and source evaluation, or explore the role of additional reader attributes needed to make inferences about information in complicated globally networked social media contexts.

Finally, by broadening our conceptualization of context beyond text and source attributes to explore the effect of one contextual attribute of the participating readers (i.e., crowdsourcing platform membership), we considered the idea that significant differences in credibility evaluations may potentially reflect the different demographic compositions of the participants in the Prolific and Toloka platforms. Unfortunately, because many different nationalities were represented within each platform, it is impossible to know more about how geographic location or nationality might have affected the credibility evaluations of participating crowdworkers in our study. However, the idea of including nationality, and other demographic attributes of readers (i.e., such as gender, ethnicity, and level of education) in future conceptualizations of credibility frameworks seems worthy of additional study.

Findings from this study also have methodological implications for research in credibility evaluation. Our full factorial design with five independent variables, each having multiple levels, allowed us to study numerous factors within the same experiment, constrained only by the number of unique variations that can be generated. Our methodology also enabled a comprehensive comparison of the influence that each considered factor had on the dependent variable. By employing a cumulative link mixed model (special form of multiple regression) to assess the impact of individual categories within each factor, the resulting coefficients and calculated effect sizes offer a means to rank these factors based on their relative impact. Our study also underscores the potential of applying software to generate and ensure that all unique combinations of social media posts have been created. Key to this approach was the well-established research in combinatorial software testing (R. Kuhn et al., 2009), which involves the automatic generation of unique test cases from a model of all factors and their levels. Crowdsourcing has emerged as a natural methodological choice in modern online combinatorial experiments that demand scalable and on-demand human subject participation. In our study, this approach enables the rapid testing of hundreds of unique variations. Additionally, leveraging multiple platforms promotes cultural and global diversity in samples, as different online human subject pools have distinct participant bases (Douglas et al., 2023). Achieving full cultural representativeness in study samples will require further efforts of course, and one must recognize the self-selection bias inherent in these subject pools. Nevertheless, platforms like Prolific, Panels & Samples (by Qualtrics), and Google Surveys are now widely recognized for capturing high-quality human opinions. They are widely adopted by multinational corporations like Meta, institutions such as The World Bank, and behavioral researchers from prestigious institutions, including Stanford and the University of Oxford.

5.2. Limitations

As with all studies, this study has limitations that must be considered when interpreting the results. First, we only investigated two first-hand evaluation strategies (i.e., discourse-based validation and knowledge-based validation), excluding corroboration from our inspection. As corroboration is a vital strategy to understand the consensus in the scientific community (Osborne & Pimentel, 2022), future studies could seek to understand how and in what conditions readers corroborate information when they encounter health information in social media. Furthermore, we did not manipulate the bandwagon features typical of social media messages, such as likes and shares, even though they have been shown to affect the credibility evaluation of such messages (Lin et al., 2016). However, we would like to note that the current design already resulted in a large number of different messages (480 unique combinations for accurate posts and 360 unique combinations for inaccurate posts).

Second, we used well-known organizations (e.g., WHO) and non-named experts (e.g., doctor, personal trainer) in the testimonial evidence. Consequently, the strength of the evidence in the testimony and research categories was partly overlapping. In future studies, the second-hand sources referred to in testimonial evidence should be better aligned.

Third, due to the platforms' functionalities, the posts were delivered to the different crowdsourcing platforms slightly differently. However, we attempted to minimize the potential effects by controlling for the platform used in our analysis. However, using two crowdsourcing platforms resulted in a more diverse sample compared to using only one platform. It is worth mentioning that the crowdsourcing platform had an effect on perceived credibility. The crowdworkers of Toloka tended to evaluate the credibility of the inaccurate posts, in particular, higher than the crowdworkers of Prolific. Future studies should consider potential effects when using different crowdsourcing platforms, especially when generalizing findings.

Fourth, using crowd-sourcing platforms for data collection has some downsides. For example, workers may complete the tasks as quickly as possible to maximize their income which is a potential threat to the validity of all studies performed using crowd-sourcing (Gadiraju et al., 2015). We combated this threat by using filters in both crowd-sourcing platforms, that is, recruiting only the top-performing workers for our study.

Fifth, we used a single item to measure the perceived credibility of the social media posts, limiting the estimation of instrument reliability. However, according to Fuchs and Diamantopoulos (2009), single-item scale may be advantageous to avoid participants resenting their being asked to respond to similar items, especially when items are administered multiple times.

6. Conclusions

Our study sheds light on how variations of multiple factors of social media posts (i.e., content, evidence type, source expertise, gender, and ethnicity) can influence credibility evaluation performance. When evaluating the credibility of

social media posts, readers can pay attention to the quality of content and trustworthiness of the source of the post. This study shows that readers tend to judge belief-consistent information as more credible than belief-inconsistent information. At worst, this tendency, accompanied by algorithm bias, may strengthen some readers' false beliefs. This, in turn, may accelerate the spread of misinformation in social media. Alarmingly, readers tend to ignore, to some extent, the quality of evidence in their credibility evaluation. This may be due to reluctance to revise one's beliefs, inability to differentiate the quality of different types of evidence, or not valuing research or expert evidence. However, regardless of whether the post was accurate or inaccurate, content posted by an expert was evaluated as more credible than content posted by a layperson. Thus, readers can be more vulnerable to misinformation posted by someone considered an expert. Moreover, our study revealed that different demographic backgrounds of the readers may significantly affect credibility evaluation performance.

Overall, our study suggests a need for further developing the credibility evaluation skills of readers across all ages, not only of children and young people who have typically been the target group of critical reading interventions. An inclusive education system should equip all citizens with adequate credibility evaluation skills. Further, the design and findings of our study answers calls to capture more of the complexities of digital reading tasks associated with texts, readers, and contexts (Coiro, 2021) in addition to calls for more credibility evaluation research that considers the demographic attributes of readers (see S. M. Shariff, 2020). Last, our study bridges theory and research efforts in psychology, communication, and media studies with those that explore critical online resource evaluation (Forzani, 2020; Forzani et al., 2022) in educational contexts.

References

Abendroth, J., & Richter, T. (2021). Mere plausibility enhances comprehension: The role of plausibility in comprehending an unfamiliar scientific debate. *Journal of Educational Psychology*, *113*(7), 1304–1322. https://doi.org//10.1037/edu0000651

Addawood, A., & Bashir, M. (2016). "What is your evidence?" A study of controversial topics on social media. *In Reed, C.(Ed.), Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 1–11. Association for Computational Linquistics. https://doi.org/10.18653/v1/W16-2801

Andreassen, R., & Strømsø, H. I. (2012). Reading about health risks: Who and what to trust? A research review. In K. P. Knutsen, S. Kvam, P. Langemeyer, A. Parianou, & K. Solfjeld (Eds.), *Narratives of risk: Interdisciplinary studies* (pp. 255–274). Waxman.

Armstrong, C. L., & McAdams, M. J. (2009). Blogs of information: How gender cues and individual motivations influence perceptions of credibility. *Journal of Computer-Mediated Communication*, *14*(3), 435–456. https://doi.org/10.1111/j.1083-6101.2009.01448.x

Barzilai, S., Thomm, E., & Shlomi-Elooz, T. (2020). Dealing with disagreement: The roles of topic familiarity and disagreement explanation in evaluation of conflicting expert claims and sources. *Learning and Instruction*, *69*, Article 101367. https://doi.org/10.1016/j.learninstruc.2020.101367

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), Article 2815. https://doi.org/10.21105/joss.02815

Braasch, J. L., Bråten, I., Britt, M. A., Steffens, B., & Strømsø, H. I. (2014). Sensitivity to inaccurate argumentation in health news articles: Potential contributions of readers' topic and epistemic beliefs. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 117–137). The MIT Press.

Brandt, E. (2022). *How ai and crowdsourcing help social scientists sample diverse populations* [Retrieved September 12, 2023]. https://venturebeat.com/business/how-ai-and-crowdsourcing-help-social-scientists-sample-diverse-populations/

Bråten, I., McCrudden, M. T., Stang Lund, E., Brante, E. W., & Strømsø, H. I. (2018). Task-oriented learning with multiple documents: Effects of topic familiarity, author expertise, and content relevance on document selection, processing, and use. *Reading Research Quarterly*, *53*(3), 345–365. https://doi.org/10.1002/rrq.197

Britt, M. A., Richter, T., & Rouet, J.-F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist*, *49*(2), 104–122. https://doi.org/10.1080/00461520.2014.916217

Chapkovski, P. (2023). Conducting interactive experiments on Toloka. *Journal of Behavioral and Experimental Finance*, *37*, 100790. https://doi.org/10.1016/j.jbef.2023.100790

Chinn, C. A., Rinehart, R. W., & Buckland, L. A. (2014). Epistemic cognition and evaluating information: Applying the air model of epistemic cognition. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 425–454). The MIT Press.

Chinn, S., & Weeks, B. E. (2021). Effects of competing statistical and testimonial evidence in debates about science. *Environmental Communication*, *15*(3), 353–368. https://doi.org/https://doi.org/10.1080/17524032.2020.1837900

Christensen, R. H. B. (2015). Ordinal—regression models for ordinal data. *R package version*, *28*, 2015.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Coiro, J. (2021). Toward a multifaceted heuristic of digital reading to inform assessment, research, practice, and policy. *Reading Research Quarterly*, *56*(1), 9–31. https://doi.org/10.1002/rrq.302

Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(1), 165–185. https://doi.org/10.1111/j.1467-9868.2004.00438.x

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of Mechanical Turk workers. *Proceedings of the eleventh ACM international conference on web search and data mining*, 135–143. ACM. https://doi.org/10.1145/3159652.3159661

Djafarova, E., & Rushworth, C. (2017). Exploring the credibility of online celebrities' instagram profiles in influencing the purchase decisions of young female users. *Computers in Human Behavior*, *68*, 1–7. https://doi.org/10.1016/j.chb.2016.11.009

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos One*, *18*(3), e0279720. https://doi.org/10.1371/journal.pone.0279720

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, *38*(1), 39–72. https://doi.org/10.1080/03057260208560187

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Feng, Y., Chen, H., & Kong, Q. (2021). An expert with whom i can identify: The role of narratives in influencer marketing. *International Journal of Advertising*, *40*(7), 972–993. https://doi.org/10.1080/02650487.2020.1824751

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). Russell Sage Foundation.

Fogg, B. J. (2002). Persuasive technology: Using computers to change what we think and do. *Ubiquity*, *2002*. Article 5. https://doi.org/10.1145/764008.763957

Forzani, E. (2020). A three-tiered framework for proactive critical evaluation during online inquiry. *Journal of Adolescent & Adult Literacy*, *63*(4), 401–414. https://doi.org/10.1002/jaal.1004

Forzani, E., Corrigan, J., & Kiili, C. (2022). What does more and less effective internet evaluation entail?: Investigating readers' credibility judgments across content, source, and context. *Computers in Human Behavior*, *135*, Article 107359. https://doi.org/10.1016/j.chb.2022.107359

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft*, *69*(2), 195–210.

Fullerton, A. S., & Anderson, K. F. (2023). Ordered regression models: A tutorial. *Prevention Science*, *24*, 431–443. https://doi.org/10.1007/s11121-021-01302-y

Fullerton, A. S., & Xu, J. (2016). *Ordered regression models: Parallel, partial, and non-parallel alternatives*. CRC Press.

Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1631–1640. ACM. https://doi.org/10.1145/2702123.2702443

Groggel, A., Nilizadeh, S., Ahn, Y.-Y., Kapadia, A., & Rojas, F. (2019). Race and the beauty premium: Mechanical Turk workers' evaluations of Twitter accounts. *Information, Communication & Society*, *22*(5), 709–716. https://doi.org/10.1080/1369118X.2018.1543443

Hämäläinen, E. K., Kiili, C., Räikkönen, E., & Marttunen, M. (2021). Students' abilities to evaluate the credibility of online texts: The role of internet-specific epistemic justifications. *Journal of Computer Assisted Learning*, *37*(5), 1409–1422. https://doi.org/10.1111/jcal.12580

Hanimann, A., Heimann, A., Hellmueller, L., & Trilling, D. (2023). Believing in credibility measures: Reviewing credibility measures in media research from 1951 to 2018. *International Journal of Communication*, *17*, 214–235. https://openaccess.city.ac.uk/id/eprint/29616

Hendriks, F., Kienhues, D., & Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: The muenster epistemic trustworthiness inventory (meti). *PloS One*, *10*(10), e0139309. https://doi.org/10.1371/journal.pone.0139309

Hervé, M., & Hervé, M. M. (2020). Package 'rvaidememoire'. *See https://CRANR-projectorg/package= RVAideMemoire*.

Hinyard, L. J., & Kreuter, M. W. (2007). Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior*, *34*(5), 777–792. https://doi.org/10.1177/1090198106291963

Hirvonen, N., Tirroniemi, A., & Kortelainen, T. (2019). The cognitive authority of user-generated health information in an online forum for girls and young women. *Journal of Documentation*, *75*(1), 78–98. https://doi.org/10.1108/JD-05-2018-0083

Hornikx, J. (2005). A review of experimental research on the relative persuasiveness of anecdotal, statistical, causal, and expert evidence. *Studies in Communication Sciences*, *5*(1), 205–216.

Hossain, M., & Kauranen, I. (2015). Crowdsourcing: A comprehensive literature review. *Strategic Outsourcing: An International Journal*, *8*(1), 2–22. https://doi.org/10.1108/SO-12-2014-0029

Jacobsen, R., Halvorsen, A.-L., Frasier, A. S., Schmitt, A., Crocco, M., & Segall, A. (2018). Thinking deeply, thinking emotionally: How high school students make sense of evidence. *Theory & Research in Social Education*, *46*(2), 232–276. https://doi.org/10.1080/00933104.2018.1425170

Karimi, M. N., & Richter, T. (2023). Thinking dispositions moderate adolescent readers' mental models of conflicting sport information. *Thinking Skills and Creativity*, Article 101233. https://doi.org/10.1016/j.tsc.2023.101233

Khosrowjerdi, M. (2020). National culture and trust in online health information. *Journal of Librarianship and Information Science*, *52*(2), 509–528. https://doi.org/10.1177/0961000619836716

Knuuti, J. (2020). *Kauppatavarana terveys–selviydy terveysväitteiden viidakossa*. Minerva Kustannus Oy.

Kuhn, D. (1991). *The skills of argument*. Cambridge University Press.

Kuhn, R., Kacker, R., Lei, Y., & Hunter, J. (2009). Combinatorial software testing. *Computer*, *42*(8), 94–96. https://doi.org/10.1109/MC.2009.253

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication*, *50*(1), 46–70. https://doi.org/10.1111/j.1460-2466.2000.tb02833.x

Larson, A. A., Britt, M. A., & Kurby, C. A. (2009). Improving students' evaluation of informal arguments. *The Journal of Experimental Education*, *77*(4), 339–366. https://doi.org/10.3200/JEXE.77.4.339-366

Lin, X., Spence, P. R., & Lachlan, K. A. (2016). Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior*, *63*, 264–271. https://doi.org/10.1016/j.chb.2016.05.002

List, A., Du, H., & Lyu, B. (2021). Examining undergraduates' text-based evidence identification, evaluation, and use. *Reading and Writing*, *35*, 1059–1089. https://doi.org/10.1007/s11145-021-10219-5

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

McCrudden, M. T., & Barnes, A. (2016). Differences in student reasoning about belief-relevant arguments: A mixed methods study. *Metacognition and Learning*, *11*(3), 275–303. https://doi.org/10.1007/s11409-015-9148-0

Meinert, J., & Krämer, N. C. (2022). How the expertise heuristic accelerates decision-making and credibility judgments in social media by means of effort reduction. *Plos One*, *17*(3), e0264428. https://doi.org/10.1371/journal.pone.0264428

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, *59*, 210–220. https://doi.org/10.1016/j.pragma.2013.07.012

Metzger, M. J., & Flanagin, A. J. (2015). Psychological approaches to credibility assessment online. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology* (pp. 445–466). Wiley Online Library. https://doi.org/10.1002/9781118426456.ch20

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Oehm, A. W., Merle, R., Tautenhahn, A., Jensen, K. C., Mueller, K.-E., Feist, M., & Zablotski, Y. (2022). Identifying cow–level factors and farm characteristics associated with locomotion scores in dairy cows using cumulative link mixed models. *PloS One*, *17*(1), e0263294. https://doi.org/10.1371/journal.pone.0263294

Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of Advertising*, *19*(3), 39–52. https://doi.org/10.1080/00913367.1990.10673191

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*(5977), 463–466. https://doi.org/10.1126/science.1183944

Osborne, J., & Pimentel, D. (2022). Science, misinformation, and the role of education. *Science*, *378*(6617), 246–248. https://doi.org/10.1126/science.abq8093

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peng, W., Lim, S., & Meng, J. (2022). Persuasive strategies in online health misinformation: A systematic review. *Information, Communication & Society*, *26*(11), 2131–2148. https://doi.org/10.1080/1369118X.2022.2085615

PewResearchCenter. (2015). *Pew research center demographic questions web or mail mode 12-29-2015* [Retrieved August 8, 2023]. https://assets.pewresearch.org/wp-content/uploads/sites/12/2015/03/Demographic-Questions-Web-and-Mail-English-3-20-2015.pdf

Prolific. (2023a). *Prolific audience checker tool* [Retrieved August 8, 2023]. Retrieved%20September%2014,%202023,%20from%20https://app.prolific.co/audience-checker

Prolific. (2023b). *Prolific participants* [Retrieved August 14, 2023]. https://researcher-help.prolific.co/hc/en-gb/categories/360000859913--Prolific-s-participants

Rezapour, M., & Ksaibati, K. (2021). Modeling crashes involving children, finite mixture cumulative link mixed model. *International Journal of Injury Control and Safety Promotion*, *28*(4), 494–502. https://doi.org/10.1080/17457300.2021.1964088

Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A two-step model of validation. *Educational Psychologist*, *52*(3), 148–166. https://doi.org/10.1080/00461520.2017.1322968

Rolin, K. (2020). Trust in science. In J. Simon (Ed.), *The routledge handbook of trust and philosophy* (pp. 354–366). Routledge.

Scharrer, L., Rupieper, Y., Stadtler, M., & Bromme, R. (2017). When science becomes too easy: Science popularization inclines laypeople to underrate their dependence on experts. *Public Understanding of Science*, *26*(8), 1003–1018. https://doi.org/10.1177/0963662516680311

Shariff, S., Sanderson, M., & Zhang, X. (2016). Correlation analysis of reader's demographics and tweet credibility perception. In N. Ferro et al. (Eds.), *Advances in Information retrieval. ECIR 2016. Lecture Notes in Computer Science, 9626*. https://doi.org/10.1007/978-3-319-30671-1_33

Shariff, S. M. (2020). A review on credibility perception of online information. *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 1–7. https://doi.org/10.1109/IMCOM48794.2020.9001724

Shaughnessy, J., Zechmeister, E., & Zechmeister, J. (2012). Repeated measures designs. In *Research methods in psychology* (9th ed., pp. 244–268). McGraw-Hill Higher Education.

Shieber, J. (2015). *Testimony: A philosophical introduction*. Routledge.

Song, H., Omori, K., Kim, J., Tenzek, K. E., Hawkins, J. M., Lin, W.-Y., Kim, Y.-C., & Jung, J.-Y. (2016). Trusting social media as a source of health information: Online surveys comparing the united states, korea, and hong kong. *Journal of Medical Internet Research*, *18*(3), e25. https://doi.org/10.2196/jmir.4193

Spence, P. R., Lachlan, K. A., Spates, S. A., & Lin, X. (2013). Intercultural differences in responses to health messages on social media from spokespeople with varying levels of ethnic identity. *Computers in Human Behavior*, *29*(3), 1255–1259. https://doi.org/10.1016/j.chb.2012.12.013

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stadtler, M., & Bromme, R. (2014, January). The content–source integration model: A taxonomic description of how readers comprehend conflicting scientific information. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 379–402). The MIT Press.

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research*, *23*(1), e17187. https://doi.org/10.2196/17187

Sun, Y., Zhang, Y., Gwizdka, J., Trace, C. B., et al. (2019). Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators. *Journal of Medical Internet Research*, *21*(5), e12522. https://doi.org/10.2196/12522

Takao, A. Y., & Kelly, G. J. (2003). Assessment of evidence in university students' scientific writing. *Science & Education*, *12*, 341–363. https://doi.org/10.1023/A:1024450509847

Toloka. (2023). *Diversity and opportunity in the crowd: 2023 global survey of tolokers* [Retrieved September 14, 2023]. https://toloka.ai/blog/global-survey-2023/

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Tufekci, Z. (2014). Social movements and governments in the digital age: Evaluating a complex landscape. *Journal of International Affairs*, *68*(1), 1–18. http://www.jstor.org/stable/24461703

van Strien, J. L., Kammerer, Y., Brand-Gruwel, S., & Boshuizen, H. P. (2016). How attitude strength biases information processing and evaluation on the web. *Computers in Human Behavior*, *60*, 245–252. https://doi.org/10.1016/j.chb.2016.02.057

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap955

Weibel, D., Wissmath, B., & Groner, R. (2008). How gender and age affect newscasters' credibility—an investigation in switzerland. *Journal of Broadcasting & Electronic Media*, *52*(3), 466–484. https://doi.org/10.1080/08838150802205801

Wertgen, A. G., & Richter, T. (2023). Source credibility and plausibility are considered in the validation of textual information: Evidence from a social media context. *Journal of Cognitive Psychology*, *35*(2), 183–204. https://doi.org/10.1080/20445911.2022.2149757

Yang, J., Counts, S., Morris, M. R., & Hoff, A. (2013). Microblog credibility perceptions: Comparing the USA and China. *Proceedings of the 2013 conference on Computer Supported Cooperative Work*, 575–586. https://doi.org/10.1145/2441776.2441841

Zarefsky, D. (2019). *The practice of argumentation: Effective reasoning in communication*. Cambridge University Press.

Zebregs, S., van den Putte, B., Neijens, P., & de Graaf, A. (2015). The differential impact of statistical and narrative evidence on beliefs, attitude, and intention: A meta-analysis. *Health Communication*, *30*(3), 282–289. https://doi.org/10.1080/10410236.2013.842528

7. Appendix

Figure 5: Pictures of the faces used in our study.

	Professor	Nurse	Professional lifestyle blogger	Parent
Asian Female				
Black Female				
White Female				
Asian Male				
Black Male				
White Male				

Table 7
Second-hand Sources for Posts Presenting Research-based Evidence

Topic	Second hand source
Fish oil	Sinn, N., Milte, C. M., Street, S. J., Buckley, J. D., Coates, A. M., Petkov, J., & Howe, P. R. (2012). Effects of n-3 fatty acids, EPA v. DHA, on depressive symptoms, quality of life, memory and executive function in older adults with mild cognitive impairment: a 6-month randomised controlled trial. British Journal of Nutrition, 107(11), 1682–1693.
Food healthiness	Shan, Z., Guo, Y., Hu, F. B., Liu, L., & Qi, Q. (2020). Association of low-carbohydrate and low-fat diets with mortality among US adults. JAMA Internal Medicine, 180(4), 513–523.
Red processed meat	Zhong, V. W., Van Horn, L., Greenland, P., Carnethon, M. R., Ning, H., Wilkins, J. T., ... & Allen, N. B. (2020). Associations of processed meat, unprocessed red meat, poultry, or fish intake with incident cardiovascular disease and all-cause mortality. JAMA Internal Medicine, 180(4), 503–512.
Vaccines	"Gołoś, A., & Lutyńska, A. (2015). Aluminium-adjuvanted vaccines–a review of the current state of knowledge. Przegl Epidemiol, 69(4), 731–734.
Vitamin D	Manson, J. E., Cook, N. R., Lee, I. M., Christen, W., Bassuk, S. S., Mora, S., ... & Buring, J. E. (2019). Vitamin D supplements and prevention of cancer and cardiovascular disease. New England Journal of Medicine, 380(1), 33–44.