UC San Diego

UC San Diego Previously Published Works

Title

Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy

Permalink

https://escholarship.org/uc/item/71p35602

Journal

Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2)

ISSN

2573-0142

Authors

Kaufman, Robert A
Haupt, Michael Robert
Dow, Steven P

Publication Date

2022-11-07

DOI

10.1145/3555611

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, availalbe at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy

ROBERT A. KAUFMAN*, University of California, San Diego, USA
MICHAEL ROBERT HAUPT*, University of California, San Diego, USA
STEVEN P. DOW, University of California, San Diego, USA

Misinformation runs rampant on social media and has been tied to adverse health behaviors such as vaccine hesitancy. Crowdsourcing can be a means to detect and impede the spread of misinformation online. However, past studies have not deeply examined the individual characteristics—such as cognitive factors and biases—that predict crowdworker accuracy at identifying misinformation. In our study (n = 265), Amazon Mechanical Turk (MTurk) workers and university students assessed the truthfulness and sentiment of COVID-19 related tweets as well as answered several surveys on personal characteristics. Results support the viability of crowdsourcing for assessing misinformation and content stance (i.e., sentiment) related to ongoing and politically-charged topics like the COVID-19 pandemic, however, alignment with experts depends on who is in the crowd. Specifically, we find that respondents with high Cognitive Reflection Test (CRT) scores, conscientiousness, and trust in medical scientists are more aligned with experts while respondents with high Need for Cognitive Closure (NFCC) and those who lean politically conservative are less aligned with experts. We see differences between recruitment platforms as well, as our data shows university students are on average more aligned with experts than MTurk workers, most likely due to overall differences in participant characteristics on each platform. Results offer transparency into how crowd composition affects misinformation and stance assessment and have implications on future crowd recruitment and filtering practices.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**.

Additional Key Words and Phrases: Crowdsourcing; Misinformation; Social Media; Bias

ACM Reference Format:
Robert A. Kaufman, Michael Robert Haupt, and Steven P. Dow. 2022. Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 553 (November 2022), 18 pages. https://doi.org/10.1145/3555611

1 INTRODUCTION

Since the early stages of the COVID-19 pandemic, misinformation about the virus has become prevalent across social media [6]. The level of misinformation has become so widespread that the World Health Organization (WHO) declared an "infodemic" characterized by "deliberate attempts to disseminate wrong information to undermine the public health response" [65] that can lead to real-world harmful physical and mental health effects [35]. In an extreme example, violent behaviors like setting fires to telephone poles were caused by misinformed beliefs tying 5G signals to COVID-19 symptoms [32]. Beginning far before the COVID-19 pandemic, misinformation has been an ongoing

*Both authors contributed equally to this research.

Authors' addresses: Robert A. Kaufman, rokaufma@ucsd.edu, University of California, San Diego, USA; Michael Robert Haupt, mhaupt@ucsd.edu, University of California, San Diego, USA; Steven P. Dow, spdow@ucsd.edu, University of California, San Diego, USA.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2022 Copyright held by the owner/author(s).
2573-0142/2022/11-ART553
https://doi.org/10.1145/3555611

issue influencing political elections [7] and fueling early anti-vaccination movements falsely linking vaccines to autism [33].

Traditionally, fact-checking experts assess information for accuracy [36]. However, due to the mass scale of misinformation on social media platforms like Twitter, there is too much demand for experts to handle the load [16, 43]. As a result, several computational approaches have sought to use machine learning (ML) to tag social media content as potentially misinformed or misleading [2, 11, 18, 34]. The nuanced language and rapidly updated nature of informational content, however, limits the success of fully automated systems [27].

To handle the volume of fact-checking on social media, there has been a recent push to use crowdsourcing or hybrid ML-crowd methods to assess the truthfulness of informational content [5, 36, 37, 50, 51, 56]. Though these studies show potential for misinformation assessment tasks, the community lacks knowledge regarding the optimal selection of crowdworkers, platforms, and the viability of crowdsourcing for the assessment of recent information and information sentiment. Since positive sentiment towards a misinformation topic can also persuade users to engage in harmful health behaviors, it is important to classify for sentiment in addition to misinformation. For example, a post expressing a personal opinion supporting the use of hydroxychloroquine for treating COVID-19 can still mislead users even if it does not cite inaccurate evidence. Therefore, in this study we seek to provide a more nuanced understanding for how crowds can be best leveraged to assess informational content for both misinformation and sentiment.

Our specific aims are fourfold. First, we seek to investigate whether crowd workers can accurately identify misinformation related to the current COVID-19 pandemic– a rapidly-evolving and highly-politicized topic at the center of recent rampant misinformation spread. We do this by employing a Tweet coding exercise where crowd workers assess the truthfulness (contains misinformation or not) of COVID-19 tweets compared with expert judgments. Second, we seek to conduct a comprehensive evaluation of how cognitive factors related to information assessment style, attitudinal biases, and belief-based biases impact misinformation detection accuracy and sentiment assessment of crowd workers. Third, we seek to highlight the importance of recruitment platform choice by drawing comparisons between two of the most popular recruitment pools: Amazon Mechanical Turk (MTurk) and university students (in this case, from a pool called SONA). Fourth, we seek to evaluate assessment performance on a novel metric relevant for examining public response within misinformation discourse: judgments of content sentiment (pro-topic, anti-topic, neutral).

Our results show that crowdsourcing is a viable method to assess content related to rapidly-developing current topics such as the COVID-19 pandemic. Some workers are more accurate than others, however. We find several factors that predict worker accuracy in misinformation judgments: respondents high in conscientiousness, Cognitive Reflection Test (CRT) scores, and trust in medical scientists aligned more with experts while workers with high Need for Cognitive Closure (NFCC) and those who lean politically conservative aligned less. We find the same factors predict accuracy on the evaluation of tweet sentiment. Importantly, these results illuminate the need to examine deeper personal factors beyond those which may be considered standard in order to more accurately predict crowd worker accuracy on both metrics. Lastly, we find large differences in coding accuracy and factors that predict accuracy between our two samples. Specifically, our university undergraduate sample (SONA) outperformed our MTurk sample at both misinformation detection accuracy and sentiment assessment.

Taken in full, our study results can enable better recruitment and filtering of crowd workers performing fact-checking assessments of misinformation on current and evolving topics like the COVID-19 pandemic. This is important both in the context of crowdsourcing itself, and in the context of using crowds to generate training sets for hybrid ML-crowd or pure ML detection platforms. Understanding personal factors related to misinformation assessment may also help

filter out individuals who may be particularly susceptible to believing misinformation and can lead to solutions to mitigate this vulnerability in future work.

2 RELATED WORK

2.1 Politicized Misinformation Spread During COVID-19

Misinformation spread has been rampant from the very early stages of the COVID-19 pandemic and has ranged from skepticism surrounding the efficacy of preventative measures such as mask wearing and vaccine uptake to rumors surrounding the virus's origins [13]. Social media platforms have been singled out as a key contributor to the spread of misinformation due to the free-flow of under-moderated information, large user bases, and the network effects of virality [54, 63].

Misinformation related to COVID-19 is distinct from misinformation within other contexts such as flat-earth conspiracy theorists [42] in that it is highly politically charged [3, 25]. Topics related to the COVID-19 pandemic including social distancing, mask use, vaccines, and potential treatments have been highly politicized since the early stages of the pandemic [25].

Barrios and Hochberg demonstrated political differences in risk perception and compliance with public health guidance on social distancing [3]. Their findings align with other studies demonstrating political conservatism is positively correlated with COVID-19 misinformation susceptibility and spreading behaviors [5, 37, 50, 53]. More generally, Roozenbeek et al. found negative correlations between misinformation susceptibility and self-reported compliance with public health guidance, including willingness to get vaccinated [53]. Lack of compliance in public health guidance can have detrimental community and individual health impacts [35, 59].

The highly-politicized nature of COVID-19 health guidelines [25] indicates that political stance is another important dimension to test for in addition to misinformation. For instance, a tweet which expresses stances that undermine public health guidelines (e.g., a tweet stating dislike for the COVID-19 vaccine) may not contain misinformation but may still promote antisocial health behavior. Sentiment assessment has been a topic of inquiry in ML and Natural Language Processing studies [27], however, it remains under-examined in the area of crowd assessment. In the present study, our participants assess the sentiment as well as the truthfulness of tweets in order to examine to what extent their assessments align with experts.

The problem of misinformation is not only important, but increasing [65]– deepening the need for quick and accurate ways to help stem misinformation spread. Understanding the role that factors such as political biases play when leveraging crowds to detect COVID-19 misinformation is critical to creating deployable solutions to the problem. In this work, we replicate known results on the effect of political orientation on judgments as well as study more specific measures on worker attitudes and beliefs, such as trust in medical scientists, that may provide deeper insight into assessment divides.

2.2 Detecting Misinformation with Crowds

Crowdsourcing can be a powerful tool to leverage and aggregate group knowledge for information tasks at potentially large scale [14]. Crowdsourcing can be especially useful in contexts where the availability of experts to complete information-verification tasks is limited. For example, Venkatagiri et al. [61] show that online expert geolocation can be augmented using crowds, empowering crowd workers to assist in online geolocation debunking efforts.

Recent studies have focused on the viability of crowdsourcing as a means to assess the truthfulness of information online. Pinto et al. [47] proposes a hypothetical crowdsourcing process for fact-checking that uses a combination of crowdwork and expert reviews. One of the earlier studies on the use of crowds for fact-checking was done by Kriplean et al. [36], who presented a fact-checking

system on public dialogue leveraging the expert skills of librarians. Zubiaga and Ji [67] studied how crowds provide credibility assessments of tweets related to Hurricane Sandy disaster management on MTurk. They found that credibility was difficult for crowds to assess, however, more details on the author of a post may improve veracity assessment of tweets. Similarly, Maddalena et al. [41] sought to use crowds to assess the credibility of news sources, finding that alignment was decent but imperfect, suggesting that assessor background may be a moderator of accuracy.

Of particular relevance to this paper are works which have used crowds to identify misinformation of political discourse and/or public health-related content. In a recent and related study, Roitero et al. used MTurk to fact-check political statements related to COVID-19, using exploratory search methods as the means for conducting evaluations [51]. Roitero et al.'s study was built on previous work [37, 50] using the same methods to assess non-recent content from Wang et al.'s *Politifact Database* [62]. Soprano et al. delved into how scales affect crowd rating alignment with experts and suggested a multidimensional scale to be used when assessing truthfulness [56]. These methods provide a meaningful foundation for the establishment of non-expert crowd-based systems as a means to fact check and tag misinformation spread online.

Despite a growing discourse on the subject, most crowdsourcing studies on misinformation assessment to date use test stimuli that is outdated [5, 36, 37, 50, 56]. The notable exception is Roitero et al. who was the first to examine recent content related to the COVID-19 pandemic [51]. We build on this past work by using stimuli on current and ongoing public health topics related to the COVID-19 pandemic, and differentiate ourselves from Roitero et al. in the nature of our task as well the depth of assessment on factors predicting crowd worker bias. Specifically, the assessment task used by Roitero et al. required participants to use web search to verify their misinformation assessment by providing a source URL, while in our task participants made judgments based on their own prior knowledge alone. This allows our participants to make far more judgments in a shorter amount of time. Testing whether misinformation related to *recent* topics can be detected by the crowd is an important step towards establishing viability for real-world deployment; crowd workers deployed to assess misinformation on social media will provide the most value if they can tag new content *before* it can spread. We seek to assess this using our streamlined task design.

The research community also lacks knowledge about the effect of the recruitment platform itself, where differences in compensation, demographics of recruitment pools, and expectations associated with each platform may impact performance. There are a number of reasons why platforms may be chosen, such as ease of access, cost, and participant skew. This is an important area of inquiry as past studies have indicated that MTurk workers may be more diverse than American university research pools [9, 24], while others have questioned the quality of MTurk task performance [10].

In sum, we find that crowdsourcing may be a viable means to assess misinformation online, however, the research community needs a better understanding of the personal factors and platform effects which may bias evaluation of both misinformation and sentiment, particularly in the context of newly emergent informational content areas.

2.3 Factors Impacting Misinformation Assessment

Cognitive biases and personal traits may impact misinformation assessment performance in the crowdsourcing domain. Though some studies have aimed at mitigating cognitive bias through task design [17], we seek to improve misinformation assessment performance by focusing our perspective on the crowd workers themselves.

Past studies indicate that worker attributes such as political orientation [5, 37, 50, 53] may affect alignment with expert-established "ground truths". Other research within the crowdsourcing domain has shown that certain cognitive and informational assessment styles are associated with misinformation assessment accuracy. Performance on the Cognitive Reflection Test (CRT), for

example, is positively associated with analytical thinking and the ability to override incorrect intuitions in order to come to a correct answer through reflection [20]. Low CRT scores have been shown to predict crowd misalignment with experts [5, 50, 51]. While previous research utilizing crowds to identify misinformation has used CRT as its main metric for measuring cognitive style, we build on this work by evaluating the impact of other variables related to cognitive style and information assessment that have not been examined in a crowd assessment context such as Need for Cognitive Closure (NFCC) [64] and Big Five Inventory (BFI) conscientiousness [30]. Due to reported associations between NFCC and misinformation susceptibility [4, 46] and negative correlations between BFI conscientiousness and misinformation spreading behaviors [1, 38], we seek to explore these relevant cognitive factors in the context of crowdsourcing.

Need for Cognitive Closure (NFCC) [64] assesses ones desire for predictability, preference for order, and discomfort with ambiguity. While previous work shows no moderation effect between NFCC and one's inclination to believe false information after multiple exposures, also referred to as the illusory truth effect [15], other research suggests that NFCC could be related to other cognitive processes associated with misinformation susceptibility. For instance, past studies show that NFCC can magnify misinformation effects in eyewitness situations [46] and can be a mechanism driving misinformation on online social networks [4]. The latter effect was attributed to avoidance behavior of high NFCC individuals when asked to provide evidence for beliefs. Information related to rapidly-updated topics such as COVID-19 health guidance may be ambiguous and unpredictable as scientific understanding is developed and communicated. As NFCC measures an individual's comfort with ambiguity, we chose NFCC for the current study to see how this trait may affect a person's ability to accurately assess the truthfulness and stance of information related to COVID-19.

The Big Five Inventory conscientiousness scale measures the tendency to be orderly, cautious, and achievement-focused [30, 31, 48]. Previous work has shown that individuals low on this metric are more likely to share fake news [38] while those higher in conscientiousness are less likely to engage in heuristic processing that could result in misinformation spread [1]. We seek to understand how this measure relates to the success of crowd workers identifying misinformation truthfulness and sentiment, as conscientiousness may play a role in how information is evaluated during the assessment task and/or how a person assesses information in their everyday life.

In past work, cognitive abilities including numeracy skill have been shown to predict assessment alignment [53]. Similarly, education level has also been negatively correlated with susceptibility to misinformation in some studies [5, 60], but not others [53]. Political conservatism is another factor shown to be negatively correlated with detecting COVID-19 misinformation [5, 37, 50, 53], however, beliefs that are more contextually-specific to evaluating social media content related to public health topics, like trust in medical scientists, have not been explored with crowds. This variable can have implications for public health outcomes, as Roozenbeek showed that lower trust in medical scientists was related to lower compliance with COVID-19 health protocols [53]. In a similar light, Pennycook and Rand used crowds to assess trust in media sources, finding higher trust in mainstream media outlets, though trust rankings depended on political affiliation of the rater and familiarity of the source [45]. In this study, we further enumerate the role that trust plays in order to differentiate it from other factors such as political orientation. Understanding these predictive factors is a critical step necessary for implementation of crowd assessment methods that can be reliable and accurate.

In the work presented in this paper, we hope to bridge the gaps between work on misinformation susceptibility– where individuals may be passively consuming content [19, 22]– with misinformation identification tasks– where crowd workers are playing an active role in critically evaluating content. We also hope to illuminate new relationships between novel factors which have yet to be explored in crowd assessment of misinformation and sentiment.

3 METHOD

This study uses a quantitative approach for assessing misinformation coding accuracy and measuring cognitive traits of workers. The study method is broken into two parts: (1) coding (i.e. labeling) task where participants evaluate 36 tweets on various health topics related to COVID-19 for truthfulness and sentiment, and (2) survey assessing background, political affiliation, and measures of cognition among other topics. This study utilized and compared two sources of crowd workers: Amazon Mechanical Turk (MTurk) and SONA, a pool of undergraduate students. Results were analyzed to reveal patterns related to truthfulness and sentiment assessment accuracy (compared to experts) and how they relate to individual characteristics measured in the survey. More specifically, t-test mean comparisons were used to assess differences in characteristics and performance of workers between platforms. Traits most associated with top performing workers were also identified. Lastly, regression analysis was used to assess and compare effects of cognitive and personal bias factors for predicting misinformation and sentiment assessment accuracy.

3.1 Online Data Collection

Participants were recruited from MTurk (n = 132) and SONA, an undergraduate psychology, linguistics, and cognitive science research pool at a large research university in the southwestern US (n = 133). The MTurk sample has a mean age of 37.3 years old (sd = 11.4) and was 42.4% female. For SONA, the mean age was 20.9 years old (sd = 3.5) and the sample was 75.1% female. All participants remained anonymous. Both samples were given identical surveys. The sample size reported does not include respondents who failed an attention check and/or finished the survey in the 10th percentile of completion time (<4.5 minutes, median completion time = 24 minutes) to ensure quality responses. Study recruitment occurred from November to December 2021. MTurk workers were compensated based on standard survey-taking rates on the platform while SONA respondents were given study credit to fulfill course requirements.

3.2 Coding Task

Respondents were shown 36 real tweets in total related to public health topics collected from Twitter and were asked to indicate whether the tweet's sentiment towards the topic (i.e., positive, negative, or neutral sentiment), and if the tweet contains or does not contain misinformation. Figure 1 shows an example of the task. 12 of the tweets were about the COVID-19 vaccine, 12 were about the hyped and debunked usage of hydroxychloroquine to treat COVID-19 [23], and 12 were about mask wearing as a COVID-19 preventative measure. Of the 12 tweets for each topic, 4 contained misinformation, 4 expressed a positive public health sentiment, 2 expressed a negative sentiment, and 2 were neutral reporting. The presentation of tweets from all topics and categories was the same. Table 1 shows examples of tweet stimuli. The paper's authors classified the tested tweets based on misinformation categories from previous work [26, 40] and whether the tweet communicates information that is contrary to scientific consensus at the time of the study period based on expert judgment.

Tweets were further categorized as "problematic sentiment" if they did not contain misinformation but still expressed sentiment that was contrary to public health guidelines (e.g., a tweet of someone saying they do not like vaccines can still dissuade others from vaccinating even if it does not include false information). Negative sentiment towards vaccines and masks, and positive sentiment towards hydroxychloroquine were defined as problematic sentiment. In total, 12 tweets containing misinformation and 6 expressing problematic sentiment were tested in the exercise across all 3 topics.

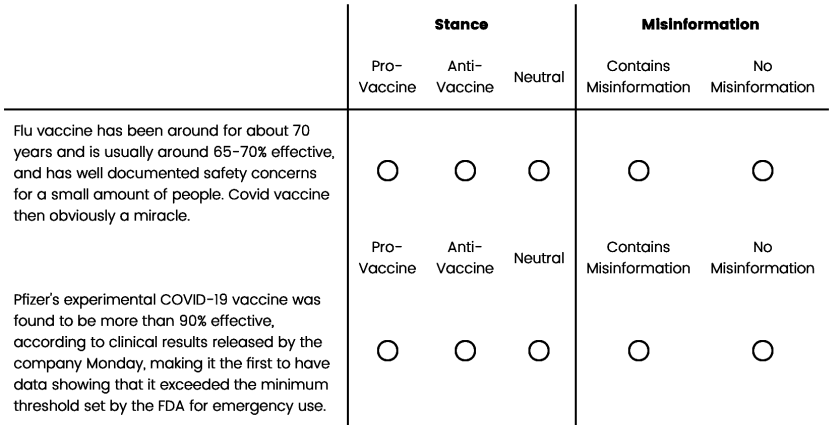

Fig. 1. Tweet Coding Task Example.

Table 1. Example tweet stimuli.

	Positive Sentiment	**Neutral**	**Negative Sentiment**
Misinfo.	Early use of Hydroxychloroquine as soon as symptoms appear reduces hospitalizations and fatalities by around 80%.Promising & it should be used.	BTW recent studies put out by hospitals show that hydroxychloroquine cures Covid.	Putting masks on children is idiotic. They inhale their own recirculated CO2 + masks don't work anyway. It's close to criminal.
Not Misinfo.	Today's vaccine news is very positive, but until large numbers can be vaccinated it is down to all of us to suppress the virus & save lives by sticking with all the rules and guidance	Many U.S. counties with low vaccination rates had a high number of positive #COVID19 tests this week. In parts of the southeast, less than 40% of people are vaccinated and more than 10% of tests were positive.	The drug doesn't work as a preventative, in the treatment of early disease and it very convincingly doesn't work in hospitalized patients... We can definitively say hydroxychloroquine doesn't work.

3.3 Survey Scales

The following survey scales were used as a part of data collection:

- *Big Five Inventory - Conscientiousness*: One 8-item subscale from the Big Five Inventory (BFI) was used to evaluate participants across the personality dimension of conscientiousness [30, 31, 48].
- *Cognitive Reflection Test (CRT)*: Three questions from the Cognitive Reflection Test were used to measure CRT. Questions from this test initially have an answer that appears "intuitive" but the correct answer requires a moment of reflection to come up with the correct answer. For example, Question 1 asks "A bat and a ball cost $1.10 in total. The bat costs $1.00 more

than the ball. How much does the ball cost?" The intuitive answer would be 10 cents, while the correct answer is 5 cents. Number of correct answers corresponds to higher CRT [20].

- *Need for Cognitive Closure*: One 15-item abridged scale [49] based on the original 42-item scale [64] used for measuring need for cognitive closure.
- *Political Orientation*: A question asking respondents to select where their political beliefs best fall under, with 1 = Very Liberal to 6 = Very Conservative. This scale is the same as those used in several prior studies showing misinformation assessment differences by political orientation [50, 51, 53].
- *Trust in Institutions*: One 4-item scale adapted from the 2019 Pew Research Center's American Trends Panel survey [21] asking respondents "How much confidence, if any, do you have in each of the following to act in the best interests of the public?". The institutions asked about were elected officials, news media, medical scientists, and religious leaders with response options ranging from 1 = No confidence at all to 4 = A great deal.

3.4 Calculating Coding Accuracy

Three scores were calculated to assess alignment between the author tweet classifications and the coding performance of the respondents. The first score called "Misinformation Correctly Labeled" sums the number of times respondents correctly labeled a misinformation tweet (highest possible score = 12). A lower score on this metric indicates higher false negative cases for identifying misinformation tweets (i.e. they did not label a tweet *with* misinformation as "containing misinformation"). The second score called "Misinformation Incorrectly Labeled" sums the number of instances respondents incorrectly indicated that a tweet contains misinformation (highest possible score = 24). A higher score on this metric indicates higher false positive rates for misinformation identification (i.e. they labeled a tweet *without* misinformation as "containing misinformation"). The third score called "Misinformation Accuracy" subtracts the Misinformation Incorrectly Labeled score from the Misinformation Correctly Labeled score. This was done to account for both types of classification errors to provide an overall metric of worker response quality (see equation below). Sentiment classification was calculated in the same manner for both positive and negative sentiment for each topic. Problematic Sentiment accuracy was calculated by summing together the accuracy scales for vaccine negative sentiment, mask negative sentiment, and hydroxychloroquine positive sentiment.

Accuracy = Correctly Labeled - Incorrectly Labeled

3.5 Categorizing Worker Performance

In order to evaluate characteristics of workers who perform the task at a standard of quality that would be helpful for research teams, we assessed differences between top performing workers and the rest. We defined top performing workers as those who labeled at least 80% of misinformation tweets correctly (10 out of 12) and did not falsely label more than 25% of non-misinformation tweets (6 out of 24). While there are multiple ways to define top performing workers, we believe 80% is an adequate cut off for correctly labeling tweets, considering in a deployed system individual ratings would likely be aggregated together. Since secondary analysis showed that workers were more likely to mislabel problematic tweets as containing misinformation compared to non-problematic tweets (57% false positive rate vs 34%), we believe 6 falsely labeled tweets would be an acceptable amount of classification error for top performing workers since it is the same number of problematic tweets in the exercise. Differences in performance and cognitive profiles were compared between the top performers and the remaining workers using t-tests.

3.6 Comparing MTurk vs SONA

To evaluate differences in coding performance between MTurk and SONA respondents, t-tests mean comparisons were conducted to detect statistically significant differences for both misinformation and sentiment categories. Multiple regression was used to compare the strength of influence between information assessment factors and bias factors for predicting crowd performance of detecting misinformation and problematic sentiment. Cognitive information assessment factors included the variables CRT, NFCC, and conscientiousness. Bias factors for the regression model were chosen based on relevance to evaluating COVID-19 misinformation, hence political orientation and trust in medical scientists were used. In order to compare effect sizes across variables, we implemented a Shapley Value regression as used in the field of economics [29, 39] and more recently in data science for interpreting machine learning models [12, 44], including models predicting COVID-19 mortality [55]. A Shapley Value regression assesses the relative importance of the predictor variables by computing all possible combinations of variables within the model and recording how much the R^2 changes with the addition or subtraction of each variable [8]. This technique allows us to determine the proportion of variance attributed to each predictor variable while controlling for potential multicollinearity.

4 RESULTS

4.1 Overall Assessment Performance

Overall accuracy across misinformation (truthfulness) and sentiment are shown in Table 2. On average, we found that participants were able to complete the task and accurately identify misinformation in over 70% of the misinformation tweets (8.57 out of 12). Almost 10 out of 24 tweets on average were incorrectly labeled for misinformation, indicating a fairly high rate of false positives. Total sample averages were high for sentiment, especially for tweets expressing problematic sentiment.

4.2 Assessment Performance Based on Recruitment Platform

4.2.1 Accuracy Differs by Platform. We find statistically significant differences between MTurk and SONA participants on the accuracy of their assessments identifying misinformation and sentiment.

As shown in Table 2, respondents recruited from SONA scored higher on average in misinformation coding accuracy (2.17) compared to MTurk respondents (-4.58). SONA respondents had a higher number of tweets correctly labeled as misinformation (9.32) and lower number of tweets incorrectly labeled (7.15) compared to MTurk respondents (7.81 and 12.57 respectively). Additionally, we find differences when assessing performance on sentiment coding, with SONA respondents on average having higher accuracy scores for positive and negative sentiment across all three public health topics compared to MTurk. SONA workers also had higher accuracy coding problematic sentiment (12.01) compared to MTurk (3.48). All of these differences are statistically significant ($p<.001$).

We note that among the respondents who were classified as top performers for misinformation coding accuracy, 27.82% were from the SONA sample compared to 8.33% of MTurkers. This difference is also statistically significant based on a chi-square test ($p<.001$).

4.2.2 Cognitive Style and Beliefs Differ by Platform. Table 3 reports differences in cognitive style and beliefs between MTurk and SONA samples. While there were no statistically significant differences in CRT between samples, MTurk respondents had higher Need for Cognitive Closure, were more conservative, and had higher trust in elected officials, news outlets, and religious leaders compared to SONA respondents ($p<.001$). SONA respondents had higher average scores in conscientiousness ($p=.002$) and higher trust in medical scientists ($p<.001$) compared to MTurk respondents.

Table 2. Mean Comparisons in Assessment Performance by Coding Type and Platform

		Overall	MTurk	SONA	Diff	*Scale Range*
	Accuracy	**-1.28**	**-4.58**	**2.17**	**6.75**	***(-24 to 12)***
Misinfo.	Correctly Labeled	8.57	7.81	9.32	1.51	*(0 to 12)*
	Incorrectly Labeled	9.85	12.57	7.15	5.42	*(0 to 24)*
	Problematic Accuracy	**7.76**	**3.48**	**12.01**	**8.53**	***(-18 to 18)***
	Vaccine Positive Acc.	0.2	-.45	.86	1.31	*(-8 to 4)*
	Vaccine Negative Acc.	2.65	1.24	4.04	2.80	*(-6 to 6)*
Sentiment	Hydroxy Positive Acc.	2.07	.84	3.29	2.45	*(-6 to 6)*
	Hydroxy Negative Acc.	0.4	-.96	1.76	2.72	*(-8 to 4)*
	Mask Positive Acc.	1.07	-.08	2.21	2.29	*(-8 to 4)*
	Mask Negative Acc.	3.05	1.40	4.68	3.28	*(-6 to 6)*

All differences between MTurk and SONA are statistically significant by T-test at **p.001

Table 3. Mean Comparisons in Factor Measures by Platform

	MTurk	SONA	Diff	p-value	*Scale Range*
CRT	.98	1.20	.22	.13	*(0 to 3)*
NFCC	4.29	3.88	.41	.001	*(1 to 6)*
Conscientiousness	3.34	3.57	.23	.002	*(1 to 5)*
Political Conservativeness	3.39	2.47	.92	.001	*(1 to 6)*
Trust Elected	2.85	2.16	.69	.001	*(1 to 4)*
Trust News	3.09	2.25	.84	.001	*(1 to 4)*
Trust Medical	3.09	3.50	.41	.001	*(1 to 4)*
Trust Religious	2.69	1.77	.92	.001	*(1 to 4)*

p-values based on T-test

4.3 Assessment Performance Based on Individual Factors

4.3.1 Factors Predicting High and Low Performers. Combining samples, we find the workers with the top performance in misinformation accuracy on average were higher in CRT and Conscientiousness, more liberal, and had higher trust in medical scientists but lower trust in elected officials, news outlets, and religious leaders compared to the remaining workers. These differences are all statistically significant, as shown in Table 4.

4.3.2 Regression Results. Shapley regression results with the combined MTurk and SONA samples (Table 5) show that CRT, political orientation, and trust in medical scientists explained the majority of the variance for misinformation coding accuracy when controlling for the other predictor variables. All predictor variables showed statistically significant effects. Based on the Beta coefficients from the multiple regression model, conscientiousness, CRT, and trust in medical scientists were positively correlated with misinformation accuracy while NFCC and political conservatism were negatively correlated.

Comparing samples, however, we find that different factors are most influential at explaining misinformation accuracy variance. For MTurk, conscientiousness was the most prominent factor explaining the majority of the misinformation accuracy variance (43.2%) followed by CRT (24.1%)

Table 4. Mean Comparisons in Factor Measures by Performer Group

	Top Performer	**Remaining Workers**	**Diff**	**p-value**	***Scale Range***
CRT	1.71	.95	.76	.001	*(0 to 3)*
NFCC	3.98	4.10	.12	.28	*(1 to 6)*
Conscientiousness	3.74	3.39	.35	.002	*(1 to 5)*
Political Conservativeness	1.96	3.15	1.19	.001	*(1 to 6)*
Trust Elected	2.19	2.57	.38	.003	*(1 to 4)*
Trust News	2.21	2.77	.56	.001	*(1 to 4)*
Trust Medical	3.81	3.17	.64	.001	*(1 to 4)*
Trust Religious	1.52	2.38	.86	.001	*(1 to 4)*

p-values based on T-test

when controlling for the other predictor variables within the model. NFCC and political conservatism were negatively correlated with misinformation accuracy while trust in medical scientists did not produce a statistically significant effect among MTurk workers and explained the lowest amount of variance.

By contrast, CRT and trust in medical scientists explained over 80% of the total variance within the SONA model (43.4% and 42.1% respectively) and were the only statistically significant variables (p<.001). Conscientiousness, NFCC, and political conservativeness were not statistically significant predictor variables among SONA respondents when controlling for the other variables within the model.

Table 5. Misinformation Accuracy: Regression Results of Cognitive and Bias Factors

		Full Sample		**MTurk**		**SONA**	
		Beta	*Shapley*	*Beta*	*Shapley*	*Beta*	*Shapley*
Cognitive Factors	Conscientiousness	1.91**	18.1%	3.15**	43.2%	.02	.8%
	NFCC	-1.41*	14.7%	-1.07*	11.8%	.46	1.0%
	CRT	1.27**	22.1%	.99*	24.1%	1.18**	43.3%
Bias Factors	Political Conservativeness	-.93**	22.5%	-.56*	13.2%	-.58	12.8%
	Trust in Med Sci	1.75**	22.6%	.79	7.7%	1.97**	42.1%
	R^2	.37		.36		.23	

Significance codes: * *p<.05,* ** *p<.001*

Table 6 shows how the same predictor variables influence accuracy for problematic sentiment. In this case, conscientiousness is positively correlated with accuracy and accounts for the most variance (29.4%) followed by political orientation at 21.9% when controlling for the other predictor variables. Political conservatism was negatively correlated with problematic sentiment accuracy. All effects in the full sample model are statistically significant (p<.001).

Breaking out by sample reveals differences in influential factors similar to our analysis of misinformation accuracy. Among MTurk workers, conscientiousness explains the majority of the variance (66.7%) followed by CRT. However, the variance is much lower for CRT (18.7%) and the effect is not statistically significant. The only other significant effect in the model besides conscientiousness is political orientation which accounts for 7.5% of variance. For SONA workers,

trust in medical scientist is positively correlated to problematic sentiment accuracy and explains the majority of variance (54.6%), followed by CRT (21.9%). Both effects are statistically significant. For both misinformation and sentiment assessment, differences related to age and sex between samples were not significant when controlling for the cognitive and bias factors.

Table 6. Problematic Sentiment Accuracy: Regression Results of Cognitive and Bias Factors

		Full Sample		**MTurk**		**SONA**	
		Beta	*Shapley*	*Beta*	*Shapley*	*Beta*	*Shapley*
Cognitive Factors	Conscientiousness	3.09**	29.4%	5.24**	66.7%	.38	3.2%
	NFCC	-1.59**	14.1%	-.65	5.3%	-.04	1.6%
	CRT	1.10**	15.6%	.70	18.7%	.87*	21.9%
Bias Factors	Political Conservativeness	-1.02**	21.9%	-.44*	7.5%	-.71	18.7%
	Trust in Med Sci	1.70**	19.0%	.23	1.8%	2.31**	54.6%
	R^2	.39		.44		.25	

Significance codes: * *p<.05,* ** *p<.001*

5 DISCUSSION

The results of this study support the viability of crowdsourcing as a method for detecting both the sentiment and truthfulness of social media content like tweets. Specifically, we study *recent* content related to the COVID-19 pandemic and show that it can be tagged using our streamlined approach.

Conscientiousness, CRT, and trust in medical scientists were positively correlated with the overall accuracy of tagging misinformation and stance while NFCC and political conservatism were negatively correlated. These effects indicate that both information assessment and bias factors influence coding accuracy of highly-politicized misinformation. The crowd recruiting platform can also impact performance outcomes: SONA respondents on average performed higher in both misinformation and sentiment coding compared to Mturk workers. Differences in cognitive style and beliefs were also detected between MTurk and SONA workers, and most likely contribute to differences in accuracy for classifying misinformation and sentiment.

These findings show that individual differences between crowd workers are important to consider when using crowds to detect misinformation, especially within the context of COVID-19. Our results imply that picking the wrong crowd could result in poor accuracy in judgment, while picking the right crowd could result in near-expert-level assessments. These results can be used to enable better recruitment and filtering practices for crowd workers assessing misinformation, and can inform the tangible goal of producing an easy-to-deploy-at-scale test of worker information veracity assessment. Though not explicitly studied, we expect our findings to generalize to related content areas, such as other public health domains.

Misinformation related to the COVID-19 pandemic presents additional challenges from previous misinformation work in that these topics are often politicized despite being public health issues. In order to accurately detect COVID-19 misinformation, specific bias factors such as trust in medical scientists and cognitive factors like CRT show high predictive value in conjunction with more general views such as political orientation. Findings such as these are particularly compelling and useful as they allow focus to move beyond umbrella-like categories like political orientation and onto measures which may be more sharply focused for an individual. The predictive value that both cognitive and biasing factors play also implies a complex relationship between abilities, attitudes, and opinions that should be explored further in future work.

Downstream implications of this work move beyond crowdsourcing as a misinformation detection method itself and onto the use of crowds to generate training sets for machine learning. Several past efforts aiming to use ML and ML-crowd paired methods to detect potential misinformation algorithmically [2, 11, 18, 27, 34]. Though the rapid spread and nuanced language of informational content presents a roadblock to such efforts [27], we highlight the potential role that crowds may play in generating training data for these models. In these cases, it is of vital importance for data generated to be accurate and unbiased in order for the ML systems to function properly, which is particularly true for "black box" systems which can be difficult to debug. In general, our streamlined methodology may be an efficient approach for generating accurate ML training sets.

5.1 Cognitive and Bias Factors Predict Accuracy

Results from the current study show that cognitive factors relating to information assessment are correlated with performance accuracy. Higher conscientiousness may be positively correlated with detecting misinformation because the trait is associated with caution and orderliness, which could lead to a higher level of thoroughness when reviewing the tweets that would be well suited for the coding task and could reflect how an individual assesses information in their everyday life. The positive correlation between CRT and misinformation detection accuracy is consistent with previous studies [5, 50, 51]. Those with higher CRT scores are more likely to have analytic thinking styles and less likely to act impulsively when making judgments, which would translate to higher performance on classification tasks such as the one used in this study. Higher NFCC was negatively correlated with coding performance. The higher need for more defined "black and white" categories might be an underlying cause for decreased performance in exercises that require the evaluation of open text posts. However, this effect was not consistent across samples and the effect size was more modest compared to the other cognitive factors.

Among bias factors, trust in medical scientists was positively correlated with increased performance. Those who trust medical scientists are more likely to agree with public health guidelines and scientific consensus, which would improve ability to detect health-related misinformation. Political conservatism was shown to be negatively correlated with detecting misinformation, and this effect is consistent with findings from previous studies. This effect may be driven by media consumption habits, where conservative media tends to downplay COVID-19 severity and public health measures [52, 57]. Consistent exposure to narratives that undermine trust in medical guidelines can influence perception of conservative workers when assessing tweets for misinformation.

Overall, the findings from this study show that dispositional traits other than CRT such as BFI conscientiousness and NFCC can also be influential factors on performance for detecting misinformation. While our results showing the impact of political orientation on misinformation detection is consistent with past literature, our findings demonstrate that more contextually-specific beliefs, such as trust in medical scientists, are also relevant factors for consideration when assessing performance.

5.2 Comparing Platforms: University Students Are More Accurate Than MTurk Workers

The results from this study show that on average, SONA respondents performed higher in both misinformation coding accuracy and sentiment coding accuracy compared to MTurk workers. A more fine-grained assessment of misinformation coding accuracy shows that SONA respondents had a higher average of correctly labeled tweets and lower average of incorrectly labeled tweets. Comparing cognitive profiles between MTurk and SONA reveal that characteristics which differentiate SONA from MTurk respondents are similar to the traits attributed to top performing workers for detecting misinformation. For example, SONA respondents, on average, had a higher

BFI conscientiousness and higher trust in medical scientists, and lower trust in elected officials, news outlets, and religious leaders compared to MTurk respondents. SONA respondents were also more liberal than MTurkers. This pattern is consistent with the difference between top performing workers and remaining workers with the exception of CRT and NFCC. While SONA respondents also have a higher average CRT score compared to MTurk workers, this difference is not statistically significant.

Regression results show that CRT, political orientation, and trust in medical scientists are the most influential factors for predicting misinformation coding accuracy when controlling for the other tested variables. For predicting problematic sentiment accuracy, conscientiousness was the most influential variable followed by CRT. Predictor variables also differed in influence based on the sample. For MTurk respondents, conscientiousness and CRT (to a lesser extent) were the most influential factors. For SONA respondents CRT and trust in medical scientists accounted for the majority of variance within the model. Differences across samples were consistent for coding problematic sentiment, where conscientiousness and CRT were the most influential for MTurk workers while trust in medical scientists and CRT were most influential for SONA. We hypothesize that conscientiousness might be a more influential factor for MTurk workers because being orderly and cautious would be more useful attributes when completing other tasks on the platform, which typically involve an attention to detail to receive approval for the work. Trust in medical scientists may be more influential for SONA participants because they are students completing the study for course credit, which may make them feel more open to classifying posts based on their beliefs. Overall, these results indicate that the context and platform used for recruiting crowd workers influence the overall characteristics of the sample, which can impact coding performance.

5.3 Limitations and Future Work

As with any study, this one is not without its limitations. First, we highlighted several personal factors in this study which predicted the coding accuracy of individual crowd workers. Due to worker fatigue-related limitations on survey length, there are several factors which may be relevant that we were not able to explore. At the most basic level, these include demographic factors such as education level, income, and comparisons between different university pools and regions. Additional factors measuring aspects of personality may also be important predictors of coding accuracy and bias. High reward dependence, low harm avoidance, and low fear of negative evaluation (especially when paired with low cognitive abilities) [66] and narcissism [28, 58] are all positively associated with higher risk to believe misinformation and conspiracy theories. Though we do not measure these factors in our current study due to limitation on survey length, they remain important areas for future work.

It is also possible that context differences between MTurk and SONA respondents could have influenced performance, since MTurkers are on a platform where the survey is one of many different tasks they could engage with to earn money while SONA respondents are taking the study to fulfill course requirements. Future crowdsourcing work should further investigate how the context of the crowdsourcing platform itself and the incentives for completing a task influence coding performance.

In this regard, tweet content does not appear in a vacuum. In this study, we stripped tweets of metadata such as the author of the tweet in order to control for bias stemming from source association. This reduces the ecologically validity of the study, though provides a clearer understanding of the effect of the tweet content itself. A future study could investigate the effect of source association (such as comparing tweets from a well known physician or politician) with judgments of accuracy.

Next, we lack a comparison group comparing our evaluation task to longer (more "effortful") or older approaches, such as those used by Roitero et al. [51]. Such a comparison group along with

corresponding metrics such as time on task or attention would give insight into the best and most efficient ways to assess misinformation using crowds. Designing an "ideal" crowdsourcing task for misinformation using empirically-derived data is an open opportunity for future work.

Finally, though data collection via survey is standard practice and widely used by prior work [50, 51, 53], there is always the risk that participants do not answer online surveys honestly, particularly when reporting identity characteristics like political orientation. We do believe the anonymity of our data collection method minimizes this risk, as we did not collect any data that could be traced to a particular individual.

Other interesting areas of future work include investigating how factors predicting crowd alignment with experts can generalize to misinformation susceptibility and spreading behavior of online social media users who are *not* explicitly tasked with evaluating content. In the wild, users tend to form opinions on the fly from limited information or even just by reading headlines [19, 22]. Thus, comparing factors of users tasked with identifying misinformation to those who may not be would be an important area of inquiry. This work would also open the door for the design of targeted solutions focused on mitigating vulnerability and/or spreading behavior, including design studies investigating the effect of different misinformation warning labels or tags.

6 CONCLUSION

Through a crowdsourcing study (n = 265), we find that crowdsourcing is a viable method to detect both the truthfulness and the sentiment of online content, like tweets, for current, evolving, and highly politicized topics such as those related to the COVID-19 pandemic. Performance, however, depends on the crowd itself. We found that respondents high in conscientiousness, Cognitive Reflective Test (CRT) scores, and trust in medical scientists were more accurate at detecting misinformation and assessing content sentiment while workers with high Need for Cognitive Closure (NFCC) and those who lean conservative politically were less accurate. Importantly, this implies that personal characteristics can be used to predict worker accuracy, and should be considered when selecting crowd workers for tasks involving misinformation identification.

We also found significant differences in the characteristics and performance of the two worker pools chosen in this study: undergraduate students (SONA) were more accurate than MTurk workers. This implies that careful consideration must be taken when choosing an online platform to host crowdsourcing tasks for misinformation identification.

Our study results can enable better recruitment and filtering of crowd workers performing fact-checking assessments of misinformation. Results have implications not just on crowdsourcing itself, but also for groups seeking to use crowds to generate training sets for ML detection systems. Future work remains on identifying additional predictors, creating ideal misinformation crowdsourcing environments, and further generalization of insights.

ACKNOWLEDGMENTS

The authors would like to express gratitude to all of the crowd workers who participated in the task and survey. We also thank Timothy Mackey, Hui Xin Ng, Joseline Chang, and Chloe Lee for their support. Funding for this research was provided by the National Science Foundation (NSF) #2009003.

REFERENCES

[1] Irum Alvi and Niraja Saraswat. 2020. Information processing—Heuristic vs. systematic and susceptibility of sharing covid 19-related fake news on social media. *J. Content Community Commun* 12 (2020), 42–56.

[2] Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of*

Data and Information Quality (JDIQ) 11, 3 (2019), 1–27.
[3] John M Barrios and Yael Hochberg. 2020. *Risk perception through the lens of politics in the time of the covid-19 pandemic*. Technical Report. National Bureau of Economic Research.
[4] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one* 10, 2 (2015), e0118093.
[5] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.
[6] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* 114, 28 (2017), 7313–7318.
[7] Ceren Budak. 2019. What happened? the spread of fake news publisher content during the 2016 us presidential election. In *The World Wide Web Conference*. 139–150.
[8] David V Budescu. 1993. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin* 114, 3 (1993), 542.
[9] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2016. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? (2016).
[10] Michael Chmielewski and Sarah C Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11, 4 (2020), 464–473.
[11] Jyoti Choudrie, Snehasish Banerjee, Ketan Kotecha, Rahee Walambe, Hema Karende, and Juhi Ameta. 2021. Machine learning techniques and older adults processing of online information and misinformation: a covid 19 study. *Computers in Human Behavior* 119 (2021), 106716.
[12] Ian Covert and Su-In Lee. 2021. Improving KernelSHAP: Practical Shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3457–3465.
[13] Jose Yunam Cuan-Baltazar, Maria José Muñoz-Perez, Carolina Robledo-Vega, Maria Fernanda Pérez-Zepeda, and Elena Soto-Vega. 2020. Misinformation of COVID-19 on the internet: infodemiology study. *JMIR public health and surveillance* 6, 2 (2020), e18444.
[14] Joseph G Davis. 2011. From crowdsourcing to crowdservicing. *IEEE Internet Computing* 15, 3 (2011), 92–94.
[15] Jonas De Keersmaecker, David Dunning, Gordon Pennycook, David G Rand, Carmen Sanchez, Christian Unkelbach, and Arne Roets. 2020. Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin* 46, 2 (2020), 204–215.
[16] Nicholas Dias and Amy Sippitt. 2020. Researching fact checking: present limitations and future opportunities. *The Political Quarterly* 91, 3 (2020), 605–613.
[17] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 48–59.
[18] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeno, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the CLEF-2019 CheckThat! Lab: automatic identification and verification of claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 301–321.
[19] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. 2018. Falling for fake news: investigating the consumption of news via social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.
[20] Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives* 19, 4 (2005), 25–42.
[21] Cary Funk, Meg Hefferon, Brian Kennedy, and Courtney Johnson. 2019. Trust and mistrust in Americans' views of scientific experts. *Pew Research Center* 2 (2019).
[22] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on Twitter?. In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*. 179–192.
[23] Joshua Geleris, Yifei Sun, Jonathan Platt, Jason Zucker, Matthew Baldwin, George Hripcsak, Angelena Labella, Daniel K Manson, Christine Kubin, R Graham Barr, et al. 2020. Observational study of hydroxychloroquine in hospitalized patients with Covid-19. *New England Journal of Medicine* 382, 25 (2020), 2411–2418.
[24] Joseph K Goodman, Cynthia E Cryder, and Amar Cheema. 2013. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26, 3 (2013), 213–224.
[25] P Sol Hart, Sedona Chinn, and Stuart Soroka. 2020. <? covid19?> politicization and polarization in COVID-19 news coverage. *Science Communication* 42, 5 (2020), 679–697.
[26] Michael Robert Haupt, Jiawei Li, and Tim K Mackey. 2021. Identifying and characterizing scientific authority-related misinformation discourse about hydroxychloroquine on twitter using unsupervised machine learning. *Big Data & Society* 8, 1 (2021), 20539517211013843.

[27] Tamanna Hossain. 2021. *COVIDLies: Detecting COVID-19 misinformation on social media.* Ph. D. Dissertation. University of California, Irvine.

[28] Sara Hughes and Laura Machan. 2021. It's a conspiracy: Covid-19 conspiracies link to psychopathy, Machiavellianism and collective narcissism. *Personality and individual differences* 171 (2021), 110559.

[29] Osnat Israeli. 2007. A Shapley-based decomposition of the R-square of a linear regression. *The Journal of Economic Inequality* 5, 2 (2007), 199–212.

[30] Oliver P John, EM Donahue, and R_L Kentle. 1991. The big five inventory: Versions 4a and 54 [Technical Report]. *Berkeley: University of California, Institute of Personality and Social Research* (1991).

[31] Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. (2008).

[32] Daniel Jolley and Jenny L Paterson. 2020. Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British journal of social psychology* 59, 3 (2020), 628–640.

[33] Anna Kata. 2012. Anti-vaccine activists, Web 2.0, and the postmodern paradigm–An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine* 30, 25 (2012), 3778–3789.

[34] Jooyeon Kim, Dongkwan Kim, and Alice Oh. 2019. Homogeneity-based transmissive process to model true and false news in social networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 348–356.

[35] Jihie Kim and Jaebong Yoo. 2012. Role of sentiment in message propagation: Reply vs. retweet behavior in political communication. In *2012 International Conference on Social Informatics*. IEEE, 131–136.

[36] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1188–1199.

[37] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. *Advances in Information Retrieval* 12036 (2020), 207.

[38] M Asher Lawson and Hemant Kakkar. 2021. Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. *Journal of Experimental Psychology: General* (2021).

[39] Stan Lipovetsky and Michael Conklin. 2001. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry* 17, 4 (2001), 319–330.

[40] Tim K Mackey, Vidya Purushothaman, Michael Haupt, Matthew C Nali, and Jiawei Li. 2021. Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter. *The Lancet Digital Health* 3, 2 (2021), e72–e75.

[41] Eddy Maddalena, Davide Ceolin, and Stefano Mizzaro. 2018. Multidimensional News Quality: A Comparison of Crowdsourcing and Nichesourcing.. In *CIKM Workshops*.

[42] Shaheed N Mohammed. 2019. Conspiracy theories and flat-earth videos on YouTube. *The Journal of Social Media in Society* 8, 2 (2019), 84–102.

[43] Salman Bin Naeem and Rubina Bhatti. 2020. The Covid-19 'infodemic': a new front for information professionals. *Health Information & Libraries Journal* 37, 3 (2020), 233–239.

[44] Ramin Okhrati and Aldo Lipani. 2021. A multilinear sampling algorithm to estimate shapley values. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 7992–7999.

[45] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[46] Gennaro Pica, Antonio Pierro, Jocelyn J Bélanger, and Arie W Kruglanski. 2014. The role of need for cognitive closure in retrieval-induced forgetting and misinformation effects in eyewitness memory. *Social Cognition* 32, 4 (2014), 337–359.

[47] Marcos Rodrigues Pinto, Yuri Oliveira de Lima, Carlos Eduardo Barbosa, and Jano Moreira de Souza. 2019. Towards fact-checking through crowdsourcing. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 494–499.

[48] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality* 41, 1 (2007), 203–212.

[49] Arne Roets and Alain Van Hiel. 2011. Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences* 50, 1 (2011), 90–94.

[50] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 439–448.

[51] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020. The covid-19 infodemic: Can the crowd judge recent misinformation objectively?. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1305–1314.

[52] Daniel Romer and Kathleen Hall Jamieson. 2021. Conspiratorial thinking, selective exposure to conservative media, and response to COVID-19 in the US. *Social Science & Medicine* 291 (2021), 114480.
[53] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* 7, 10 (2020), 201199.
[54] Hans Rosenberg, Shahbaz Syed, and Salim Rezaie. 2020. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Canadian journal of emergency medicine* 22, 4 (2020), 418–421.
[55] Matthew Smith and Francisco Alvarez. 2021. Identifying mortality factors from Machine Learning using Shapley values–a case of COVID19. *Expert Systems with Applications* 176 (2021), 114832.
[56] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management* 58, 6 (2021), 102710.
[57] Dominik A Stecula and Mark Pickup. 2021. How populism and conservative media fuel conspiracy beliefs about COVID-19 and what it means for COVID-19 behaviors. *Research & Politics* 8, 1 (2021), 2053168021993979.
[58] Anni Sternisko, Aleksandra Cichocka, Aleksandra Cislak, and Jay J Van Bavel. 2020. Collective narcissism predicts the belief and dissemination of conspiracy theories during the COVID-19 pandemic. (2020).
[59] Linda Thunström, Stephen C Newbold, David Finnoff, Madison Ashworth, and Jason F Shogren. 2020. The benefits and costs of using social distancing to flatten the curve for COVID-19. *Journal of Benefit-Cost Analysis* 11, 2 (2020), 179–195.
[60] Jan-Willem van Prooijen. 2017. Why education predicts decreased belief in conspiracy theories. *Applied cognitive psychology* 31, 1 (2017), 50–58.
[61] Sukrit Venkatagiri, Jacob Thebault-Spieker, Rachel Kohler, John Purviance, Rifat Sabbir Mansur, and Kurt Luther. 2019. GroundTruth: Augmenting expert image geolocation with crowdsourcing and shared representations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[62] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).
[63] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine* 240 (2019), 112552.
[64] Donna M Webster and Arie W Kruglanski. 1994. Individual differences in need for cognitive closure. *Journal of personality and social psychology* 67, 6 (1994), 1049.
[65] World Health Organization (WHO). 2020. Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation
[66] Bi Zhu, Chuansheng Chen, Elizabeth F Loftus, Chongde Lin, Qinghua He, Chunhui Chen, He Li, Robert K Moyzis, Jared Lessard, and Qi Dong. 2010. Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities. *Personality and Individual differences* 48, 8 (2010), 889–894.
[67] Arkaitz Zubiaga and Heng Ji. 2014. Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining* 4, 1 (2014), 163.

Received January 2022; revised April 2022; accepted August 2022