Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria

MD MOMEN BHUIYAN, Virginia Tech
AMY X. ZHANG, University of Washington
CONNIE MOON SEHAT, Hacks/Hackers
TANUSHREE MITRA[*], University of Washington

Misinformation about critical issues such as climate change and vaccine safety is oftentimes amplified on online social and search platforms. The crowdsourcing of content credibility assessment by laypeople has been proposed as one strategy to combat misinformation by attempting to replicate the assessments of experts at scale. In this work, we investigate news credibility assessments by crowds versus experts to understand when and how ratings between them differ. We gather a dataset of over 4,000 credibility assessments taken from 2 crowd groups—journalism students and Upwork workers—as well as 2 expert groups—journalists and scientists—on a varied set of 50 news articles related to climate science, a topic with widespread disconnect between public opinion and expert consensus. Examining the ratings, we find differences in performance due to the makeup of the *crowd*, such as rater demographics and political leaning, as well as the scope of the *tasks* that the crowd is assigned to rate, such as the genre of the article and partisanship of the publication. Finally, we find differences between expert assessments due to differing *expert criteria* that journalism versus science experts use—differences that may contribute to crowd discrepancies, but that also suggest a way to reduce the gap by designing crowd tasks tailored to specific expert criteria. From these findings, we outline future research directions to better design crowd processes that are tailored to specific crowds and types of content.

CCS Concepts: • **Human-centered computing** → *Human computer interaction (HCI)*; *Empirical studies in HCI*;

Keywords: misinformation, crowdsourcing, credibility, news, expert

ACM Reference Format:
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. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, CSCW2, Article 93 (October 2020). ACM, New York, NY. 26 pages. https://doi.org/10.1145/3415164

1 INTRODUCTION

A misinformed citizenry, when it comes to critical issues impacting public health and public policy such as climate change and vaccine safety, can lead to dangerous consequences. As misinformation proliferates online, social and search platforms have sought effective mechanisms for tackling harmful misinformation [60]. One strategy that many online platforms have deployed is partnerships

[*] This work was conducted while the author was at Virginia Tech.

Authors' addresses: Md Momen Bhuiyan, Virginia Tech, momen@vt.edu; Amy X. Zhang, University of Washington, axz@cs.uw.edu; Connie Moon Sehat, Hacks/Hackers, connie@hackshackers.com; Tanushree Mitra, University of Washington, tmitra@uw.edu.

2573-0142/2020/10-ART93 $15.00
https://doi.org/10.1145/3415164

with expert groups to judge the credibility of articles posted on their platform. Initiatives include Facebook's fact-checking program, which employs third-party groups such as Climate Feedback's community of science experts [63] to rate articles that then get a warning label or down-ranked in users' feeds [18]. However, expert feedback is hard to scale, given the relatively small number of professional fact-checkers and domain experts. Thus, in recent years, platforms and third-party organizations have developed tools and processes to relax the expertise criteria needed to judge the credibility of news articles. Initiatives that have pursued a low-barrier crowdsourced approach to fact-checking include TruthSquad [21], FactcheckEU [19], and WikiTribune [51]. However, these prior attempts have had their own issues with maintaining high quality at scale, due to crowd-sourced content requiring additional input from a relatively small number of editors [4]. As a result, most crowdsourced approaches still do not scale well, due to needing final judgments by experts or only using crowds for secondary tasks while primary research is still delegated to experts [4]. Despite the interest in scaling up news credibility assessment, there is still a great deal that is unknown about when and how crowd credibility assessments align with experts.

How can we better understand the considerations to take into account when developing scaleable crowdsourced processes for news credibility assessment? In this work, we investigate three components of crowdsourcing news credibility in particular: how crowd alignment with experts changes with regards to the background and identity of the *crowd*, the scope of *task* in terms of the news content being assessed, and the type of *expert criteria* being measured against. We gather a large dataset of 4,050 news credibility ratings, spanning 4 types of raters (2 crowd groups and 2 expert groups) and 81 individuals in total, on 50 articles in the domain of *climate science*—an area with widespread disconnect between public opinion and scientific consensus. Through a focus on climate science, a field in which strong expert consensus exists, we provide a more consistent basis upon which to hypothesize efforts to crowdsource other topics overall, including those with less expert consensus (e.g., emerging knowledge of COVID-19, politics) [1]. The crowd raters we compare include journalism students and Upwork workers. We contrast news credibility ratings from these two groups with ratings by experienced journalists and climate scientists. All data collected for this work has been released publicly[1], with individual identities anonymized.

We find that about 15 ratings from either the journalist students or the Upwork workers are needed in order to achieve 0.9 correlation with journalism experts. However, when it comes to science experts, 15 ratings from either crowd group only result in 0.7 correlation with scientists. Overall, we find little difference between our two crowd groups in terms of correlation to experts. But when we examine across crowd groups to consider how the personal traits of age, gender, educational background, and political leaning alter ratings, we find that raters with less education and those who were not Democrat have higher disagreement with experts.

Besides differences due to the makeup of the crowd, we additionally determine differences in credibility ratings by the kind of content being evaluated in the requested task. When we break down how different groups' ratings differ according to characteristics of the article, such as its genre and the partisanship of its publication, we find that crowd groups have stronger correlation with experts on opinion articles and articles from more left-leaning publications.

Finally, our analyses uncover differences in the criteria used to determine credibility between our two expert groups on their news credibility ratings. These differences flow to crowds—as science and journalism experts disagree more on a piece of news, crowd raters disagree more as well. In order to understand why experts disagree, we gather 147 open-ended explanations by our experts regarding the criteria they used to make their ratings. We find that science experts put emphasis on criteria related to accuracy, evidence, and grounding presented in the article, while

[1] Data: https://data.world/credibilitycoalition/credibility-factors2020

journalism experts stress publication reputation. This difference may explain why crowds have greater correlation with journalists rather than scientists.

The differences in expert criteria of what constitutes credibility, along with our findings on differences in crowd performance based on background and article type, suggest a future line of work to design crowdsourced news credibility processes that are tailored towards particular types of expertise. Instead of broadly rating credibility, different crowd rating tasks might align with different experts. In the case of straight reporting of climate-related conferences or events, for example, one might ask crowds to align more with signals used by journalism experts, whereas reporting on scientific conclusions might align more with signals tied to science expertise. We discuss this possibility and present some preliminary findings in our Discussion. At a high level, our results suggest two strategies—*person-oriented* and *process-oriented*—to improve task design by respectively filtering on rater background during the recruitment and training devised towards reducing particular differences. By taking into account diverse expert criteria and task fitness into these strategies, future designers may improve the reliability of crowdsourced news credibility. Altogether, our work offers a deeper understanding of the conditions under which crowdsourced annotations might serve as a proxy for different forms of reliable expert knowledge.

2 RELATED WORK

2.1 Credibility

Credibility is often defined as a multi-dimensional construct comprising believability [23], fairness [26], reliability [59], quality [66], trust [32], accuracy [22], objectivity/bias [15, 44] and "dozens of other concepts and combination thereof" [30]. Compared to other works, credibility has been defined by Flanagin and Metzger as made up of two primary dimensions: *trustworthiness* and *expertise* [20]. Oftentimes, credibility is targeted at just the message and/or the source, while some extend it to consider context, such as the channel or medium where the message is published [36, 42]. However, research has also shown that receivers often do not distinguish between message source and the medium [12]. Furthermore, scholars from information science to cognitive psychology can range in their definition of credibility as a purely objective assessment or a subjective judgment by the information receiver, adding complexity to the primary dimensions [20, 22, 57]. Despite significant scholarly work in multi-disciplinary domains, the definition of credibility and its measurement still lacks a unified strategy [30]. Consequently, in this work, we approach credibility as a blend of subjective and objective assessments of the "message," in this case, the news article.

2.2 Crowdsourcing News Credibility Assessment

Though much has been made about the "wisdom of crowds," it is still unclear whether crowdsourcing can be an effective strategy for assessing news credibility and misinformation in a reliable and systematic way. Partly this has to do with the limits of crowds on certain topics. It is accepted that collective wisdom can be better than an individual's judgment, including those of individual experts [69]. These conclusions are based upon mathematical principles, which however also indicate the converse—that in certain circumstances, the collective can perform a great deal worse.

One circumstance is when crowds do not have enough *relevant* information, suggesting that a baseline expertise is necessary [3, 68]. Crowds may also make mistakes due to an incorrect general perception about whether a piece of information is false or true [3]. Other characteristics of the crowd, such as its diversity, size, and suitability towards the task in question also play a part [40, 47, 71, 73]. Given this prior work, the question we consider then is not *whether* crowdsourcing is a viable approach for news credibility assessment but instead under *which conditions* can we unlock the "wisdom of select crowds" [39].

Prior literature suggests that some segments of the population are potentially worse at assessing news. For example, research has found that conservative-leaning, older, and highly politically-engaged individuals are more likely to interact with "fake news" in the U.S. [28, 41, 72] In addition, strong analytical thinking is associated with increased capacity to discern true headlines from false or hyperpartisan ones [58]. Certain topics can be polarizing for audiences, leading to poor alignment with experts for portions of the public with a particular political leaning, such as in the case of climate science [25]. Yet other prior work shows that laypeople even in polarized contexts are able to discern high quality content from low quality ones [53] and are overall highly correlated with ratings from professional checkers [16]. Research has also found that homogenous groups of people can help increase accuracy while reducing polarization—strengthening the case for crowdsourced ratings [6]—an aspect we delve into while focusing on credibility assessment of news articles pertaining to climate science, a highly polarized topic among non-experts.

2.3 Task Suitability of News Credibility Assessment by the Crowd

Though crowds' performance may vary depending on demography, their performance can also depend on what task is being asked of them. For example, researchers have encountered differences when the public is asked to fact-check versus assess media trustworthiness [4, 54, 61]. Because crowdsourced fact-checking continues to prove challenging, a subjective rating task like trustworthiness might be far less complex and better suited to crowds than fact-checking [4]. In fact, due to this difficulty in fact-checking, research shows that some topics (e.g., economy and politics) have higher probability of getting asked to be checked than others (e.g., education and environment) [29]. There may also be differences when it comes to the unit of content analysis: claims, tweets, articles, and sources [11, 46, 52]. Additionally, the subject area of news coverage may make a difference; some topics may be easier to understand, such as events versus specialized science or health news. Research has also found that most Americans do only slightly better than chance at distinguishing factual from opinion news statements [45], and half are unfamiliar with the term "op-ed" [24]. This is concerning as opinion pieces have different journalistic standards compared to news articles. Finally, as mentioned previously, readers' political biases may also play into their assessment of a piece of content [43]. This is why in order to assess these content-level constraints, we analyze the performance of crowds on articles divided by genre and the political leaning of the article's source.

2.4 Differences in Criteria for Expert Assessments

Finally, little is known about how *different experts* make use of the information embedded in news content in their credibility judgments. That is, many crowd assessments measure a crowd's alignment with a body of experts from a single domain, but multiple expertise can be in scope in terms of news credibility—in our case, scientific and journalistic. Thus, there might be different criteria against which an approach at scale may wish to align. For example, while examining how finance and health experts rank websites in their respective fields, scholars found some innate differences in respective domains (e.g., nature of information in one domain being "proven" versus another one being "predicted"), as well as experts' behavioral differences in perceiving website characteristics (e.g., differences in emphasis on visual characteristics) [64]. While one might try to control for such intra-domain differences among experts by careful selection of the topic (e.g., where the majority of the experts agree such as in climate science [1]), our understanding of how different domain experts would judge the same piece of news content is still limited. We fill this gap by examining the different criteria used by domain experts—in our case, environmental scientists and journalists—when it comes to credibility.

Overall, the assumption that a relationship between crowds and experts can be established in a meaningful way at scale underlies many approaches in the field, and it is the approach to this relationship that this study seeks to complicate.

3 STUDY DESIGN

In this work, we conduct an investigation into three major considerations for crowdsourcing news credibility at scale. Based on the literature thus far, we expect that the crowd and subject area experts will perceive the credibility of news information differently. To systematically and empirically understand this difference, we consider the following dimensions:

- Differences in ratings might reside in the *raters*, as some raters are likely to be more in alignment with expert judgment. Aspects about the background of these raters could perhaps help select suitable raters.
- Other differences might reside in the *task* they are given—in this case, the articles they are assigned to assess, as news articles can vary along several spectra. For example, raters and experts may differ in noteworthy ways as they evaluate opinion pieces as opposed to "straight" news, or articles that have perceptible political lean.
- Finally, differences might reside in what *criteria* is used to judge credibility in news stories. If experts are using different criteria to determine credibility, some of them may be more or less accessible to or mirrored by crowd raters.

In order to understand these potential differences, we ask the following research questions:

- RQ1: How do crowd raters compare with experts when it comes to news credibility assessments?
- RQ2: How do personal characteristics of age, education, gender, and political leaning affect credibility ratings from the crowd?
- RQ3: How do characteristics of news articles, such as article genre (news, opinion, analysis) and political lean of the publication, affect credibility ratings?
- RQ4: How do experts in science versus journalism differ in the criteria they use to assess news credibility?

The first RQ confirms the initial assumption that experts and crowd raters disagree. However, the differences are not uniform—instead, we see that crowd raters tend to agree with journalism experts more, and that as experts disagree more, crowd raters disagree more as well. Surprisingly, we find no major differences between the two populations we recruited from—journalism students and Upwork workers. We further explore in RQ2 the suitability of different segments of the crowd for assessing news credibility. We do find differences across the board based on educational background and political leaning. RQ3 then focuses on the nature of task suitability for crowd raters according to the characteristics of articles, finding that crowds correlate more with experts when it comes to opinion articles and left-leaning publications. Interestingly, journalism and science experts also correlate more closely with each other in those cases, while having greater disagreements when it comes to the news and analysis genres and center-leaning publications.

Some of these differences can be illuminated by RQ4, which delves into the criteria that different experts use to assess news credibility. We find that science experts focus more on the evidence presented in the article and the underlying accuracy of the claims, while journalism experts focus on the publication reputation of the news outlet and overall professional standards. Indeed, for some types of articles, such as ones that report on a press conference, the criteria used by journalists may be more relevant, while for other types of articles, such as ones that report on a new scientific finding, scientists' criteria may be preferred. These differences may also explain why crowds align more with journalists, as the criteria they use may be easier for crowds to assess. We conclude with

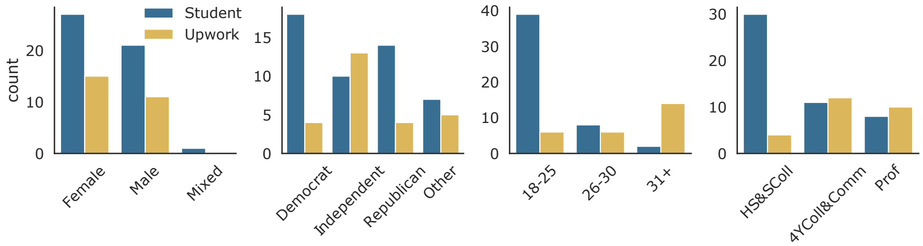


Fig. 1. The figures show user distribution by gender, political party, age and education for our crowd groups. For education, "HS&SColl", "4YColl&Comm" and "Prof" stands for respectively "High School & Some College No Degree & Some College", "4 Year College & Community College/Vocational Training" and "Professional & Graduate Degree".

a discussion of how to design news credibility assessment tasks that are tailored to specific crowds and contents.

3.1 Topic Area and Articles

We wished to isolate differences in crowdsourcing based in the raters, tasks, or between disciplinary fields themselves as opposed to disagreements due to lack of internal consensus among subject area experts on the underlying facts. For this reason, we chose to focus on scientific topics with a high degree of consensus among domain experts, as opposed to political topics in which the potential for stable ground truth is much more challenging. We also needed a subject matter that generates enough examples of news and in which misinformation or problematic information appear regularly, as these are the conditions under which major platforms are operating when surfacing articles to fact-checkers.

Thus, we selected 50 articles focusing on climate and environment issues, a topic that has a high degree of consensus among science domain experts but that has also become politicized. To gather articles, we began with the Buzzsumo social media research tool in late 2018 to find the most popular English-language articles over the previous year with the keywords of "climate change," "global warming," "environment," and "pollution." Then, among the top results, we selected a set of articles with varying amounts of scientific reference. We also sought to diversify the number of outlets publishing the articles. In addition, we sought to include a range of liberal to conservative positions or attitudes towards climate problems in the article selection[2].

We expect a certain amount of correlation between conservative positions and less credible information on climate science, based on past studies, that may not generalize to other topics. But by conducting a deeper exploration of a single domain, we gain richer evidence upon which we can make inferences regarding the reasoning behind certain differences in ratings. This allows our study to consider implications for design more broadly across the dimensions of raters, tasks, and expert criteria despite being grounded within a single domain.

3.2 Raters

We collected credibility ratings on articles from four different groups, including two crowd groups consisting of: 1) 49 participants recruited from journalism and media schools, as well as 2) 26 Upwork crowdworkers, and two expert groups comprising: 3) three climate scientists, and 4)

[2] See Appendix B for our article distribution across sources.

three journalists. Each crowd and expert rater rated all 50 articles in our dataset. Demographic information for the crowd groups can be found in Figure 1.

Students: The first group was canvassed through the *Credibility Coalition*[3] network, which has worked directly with nonprofits and journalism schools to build up a cohort motivated to combat misinformation. They are predominantly pursuing higher education in the U.S. and tend to be politically liberal. The Credibility Coalition actively recruited, e.g., with campus Republican clubs, to achieve more demographic balance for the study.

Upwork: In addition, we also used the *Upwork* platform for freelancers to gather from a more general population. For this study, we restricted participants to the U.S. Participants were admitted on a first-come basis until demographic balance became an issue (i.e., politically liberal respondents were declined once more conservatives were needed for balance).

With regards to experts, despite the realistic challenge of recruiting people with subject area expertise to participate due to their other obligations, we nevertheless sought more than a single expert's input to be able to capture how experts differ amongst themselves. In total, we recruited three experts each for the two types of expertise represented in this study.

Scientists: Three science experts were directly referred to us by contacts at major science organizations, including Climate Feedback, AAAS, and the National Academy of Sciences. Two of our experts are male, one is female. All three of our experts possess a Ph.D. in a climate-related field: two related to oceanography and atmospheric science, and one that intersects environment and economics.

Journalists: Our three journalism experts, reached through personal networks, each possess at least seven years of professional journalism experience in the U.S. Professional experience means that they received compensation for full-time positions within the journalism industry as writers, editors, and reporters of stories. Two of the experts are male, one is female. Two of our experts worked for major national newspapers while one worked for major broadcast news networks.

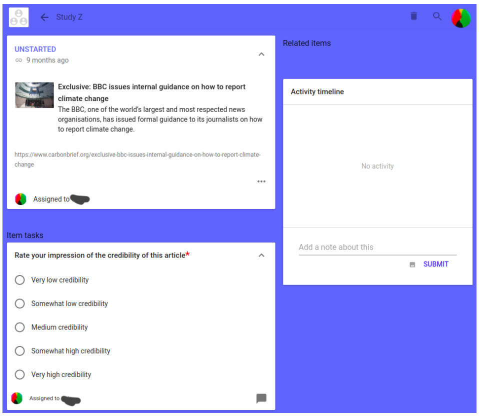
To clarify the difference between expert fields, our news experts were not science journalists. It is worth noting that science articles can be written by non-science journalists, especially amid the downsizing of news departments and as seen with sports desks writers who have recently been reassigned to coronavirus beats [27, 38]. In addition, the relationship between science experts and news professionals need not always be harmonious: sometimes journalists provide a needed function of accountability and transparency outside of the scientific profession [7].

3.3 Rater Tasks

The approach for this study kept the challenge of large-scale information assessment in mind. For this reason, the questionnaire was designed to be short, in order for raters to be able to assess many articles. Before participation, crowd raters filled out a demographic survey. We also required crowd raters to commit to an Annotator Code of Conduct provided in their informed consent, which included performing their duties in as accurate and diligent manner as possible, and avoiding conflicts of interest.

All raters, crowd and expert, received reading and rating tasks as shown in Figure 2 using an annotation platform called Check[4]. For each article, all raters were asked to read the article and provide their perception of the article's credibility on a 5-point Likert scale, ranging from *very low* (1) to *very high* (5). Crowd raters completed all tasks across a 7–10 day period (estimated at 10 hours total) with a recommended limit of 10–15 minutes per article. After completing 50 articles on time, they received the full payment of $150.

[3] https://credibilitycoalition.org
[4] https://meedan.com/check



Study Z

UNSTARTED

9 months ago

Exclusive: BBC issues internal guidance on how to report climate change

The BBC, one of the world's largest and most respected news organisations, has issued formal guidance to its journalists on how to report climate change.

https://www.carbonbrief.org/exclusive-bbc-issues-internal-guidance-on-how-to-report-climate-change

Assigned to [redacted]

Item tasks

Rate your impression of the credibility of this article*

○ Very low credibility

○ Somewhat low credibility

○ Medium credibility

○ Somewhat high credibility

○ Very high credibility

Assigned to [redacted]

Related items

Activity timeline

No activity

Add a note about this

SUBMIT

Fig. 2. An image of the questionnaire in the Check annotation tool.

In addition to providing ratings, all six expert raters optionally provided an open-ended rationale for their credibility rating for each article, resulting in 147 rationales out of a possible 300 across all experts and articles. After the completion of 50 articles, expert raters received a payment of $300.

Finally, we asked only the journalism experts to additionally classify each article across three categories: *News*, *Opinion*, and *Analysis* (understood as a close examination of a complex news event by a specialist [62]). We consulted journalism experts while developing these three categories along with the ability to select *Not Sure*. This would allow us to better understand the potential for genre-related differences in our analysis. Of the articles, 48 of them had a majority genre applied by the three experts, with 32 classified as *News*, 8 as *Opinion*, and 8 as *Analysis*.

3.4 Methodology

Much of our analysis includes inter-rater reliability, correlation between groups, and a series of regressions. Throughout, we used Krippendorf's alpha for inter-rater reliability which is appropriate for differing data types including ordinal, nominal, and interval. For correlation analysis, we used Spearman's rank correlation—a nonparametric measure of the strength and direction of association between two variables. To realize the required number of raters needed, we performed a power analysis with settings including a significance of 0.05, a large effect size of 0.5 and a power of 0.8 [9]. This resulted in a required sample size of 29. For robustness in the analysis and to account for sampling error, we calculated the correlation 100 times by bootstrapping, similar to related work [47]. Additionally, we used a general ordinary least squares (OLS) linear regression on our data. Such a regression model despite less-than-perfect fit compared to non-linear models, have greater interpretability.

4 RESULTS

4.1 RQ1: Comparing UpWork and Student Crowd Raters to Science and Journalism Experts

We begin by analyzing the credibility ratings made by our two crowd rating groups and compare their ratings with ratings made by our two expert groups. Considering all the ratings we collected from each group, Table 1 presents the inter-rater reliability (IRR) and average credibility ratings within each of our two crowd groups—`Student` and `Upwork`—and our two expert groups—`Science` and `Journalism`. Overall, we see that the experts had much higher IRR within each group than the

	#	α	Avg. Credibility Rating (Std. Dev.)
Student	49	0.44	3.49 (1.32)
Upwork	26	0.48	3.34 (1.33)
Expert[Science]	3	0.75	3.21 (1.27)
Expert[Journalism]	3	0.83	3.60 (1.42)

Table 1. Inter-rater reliability using Krippendorff's alpha (α) within all 4 rater groups on the question of credibility across 50 articles, along with average credibility rating.

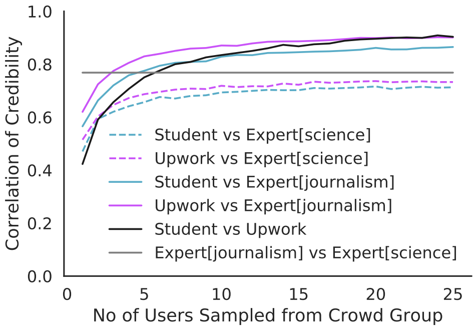


Fig. 3. Correlation of credibility ratings among all pairs in four groups: 2 crowd and 2 expert groups. In each crowd group, we sample the number of raters from 1–25. For expert groups, we take all 3 ratings. Then we compute the Spearman ρ between the mean responses from each group on all 50 articles. The plot shows average ρ after 100 resamplings.

crowd raters, with the journalists most aligned at 0.83. We also compute the correlation within each expert group, i.e., comparing one expert with the other two. Again, science experts show lower correlation (sci_1=0.72, sci_2=0.72, sci_3=0.62, $jour_1$=0.80, $jour_2$=0.77, $jour_3$=0.80). We note that our one scientist with 0.62 correlation with the other scientists comes from a social science and environmental studies background as opposed to purely environmental studies, demonstrating that specific expertise even within a field could potentially give rise to differences in credibility assessment. On average, science experts had the lowest average credibility scores while journalism experts had the highest, and the two crowd groups were in between.

We also compute the correlation of credibility ratings among all combinations of groups using Spearman's ρ. Figure 3 shows the pairwise correlation between rater groups when we vary the number of raters from 1 to 25 in `Student` or `Upwork`. We randomly sample 100 times from each group and then average the result; using this strategy, no individual rater has undue weight. This approach has also been used in prior studies for reliably comparing large crowds with limited expert ratings [46]. With only 3 raters in each group of experts, we simply average them per group. We find that the correlation between the two expert groups is 0.77. Correlation between the two crowd groups starts off low at about 0.4 with only 1 rater, but becomes high (ρ = 0.9) with about 15 or more raters within each group. This suggests that when averaging across 15 or more raters, both rater populations begin rating about equivalently. To account for lack of demographic control between the two crowd groups, we performed similar analysis on a matched data set shown in Appendix A. We find some minor differences including a slight lowering of the correlation between the two crowd groups as well as between students and experts. However, results from the matched data do not contradict our findings, offering additional confidence to our overall results.

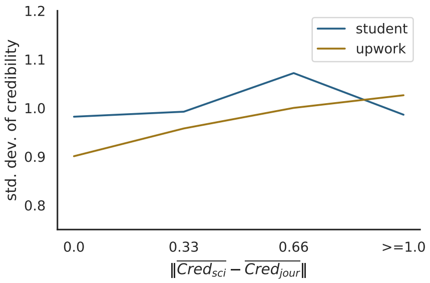


Fig. 4. Changes in standard deviation of crowd groups' credibility rating as absolute distance between two expert groups' average credibility rating grows.

4.1.1 Both Student and Upwork crowd raters are more aligned with journalists than with scientists. When we dive into the correlation of each crowd group to each expert group, differences emerge. First, we notice that Upwork has slightly higher correlation with both sets of experts than Student. The gap, while small in both cases, is nonetheless robust in the case of journalists (0.04, $t = 2.31$, $p < 0.02$) averaging across 1–25 raters. In the case of scientists, the gap was 0.02 ($t = 1.59$, $p < 0.11$). Second, we note that it takes about 15 crowd raters to achieve about 0.87 correlation with journalists. (0.85 for Student and 0.89 for Upwork). However, crowd raters get only about 0.72 (0.71 for Student and 0.73 for Upwork) correlation with scientists using 15 raters, and ratings do not improve at 25 raters. The difference between correlation with scientists versus journalists is a major one, with crowds aligning with journalists more (0.13 difference for Student and 0.15 difference for Upwork). However, recall that our analysis of correlation within individual experts show a range between 0.6 and 0.8. Both sets of crowd raters at 15 ratings each still fall within that range in their correlation with experts.

4.1.2 As science and journalism experts disagree more, crowd raters disagree more as well. Finally, we examine how our crowd groups' ratings change when expert groups diverge in their rating from each other. Figure 4 shows the plot of standard deviation of the crowd workers as the absolute difference between the average credibility ratings of the two expert groups goes up. The figure shows an almost linear upwards trend for the Upwork workers. Students also have an upward trend initially, though this trend reverses at the last point. Comparing the two crowd groups, we find medium correlation between their standard deviations (Spearman $\rho = 0.59$, p<0.001). Unsurprisingly, as the disagreement grows between the expert groups, credibility ratings of the crowd also diverges. In RQ3 and RQ4, we examine in more detail the articles that lead to higher expert disagreement, finding that factors include the type of article and differences in expert criteria regarding credibility.

4.2 RQ2: Personal Factors Affecting Credibility Ratings Among the Crowd

Next, we examine more deeply the crowd raters and consider their demographics. To determine how crowd raters' personal characteristics, such as their age and gender, relate to how well they agreed with experts, we perform an OLS regression on the error in our crowd raters' credibility rating when compared to experts' average rating. In Tables 2 and 3, we present 6 models, where ratings from just Student, just Upwork, and then Student and Upwork *combined* are compared against ratings from Science and then Journalism. We re-coded crowd raters' education responses into three larger groups due to low quantities for some of the responses: combining "High School", "Some College No Degree" and "Some College" into one and "4 Year College" with "Community

	Expert[Science]								
	Student			Upwork			Stud.+Upwork		
	β (sig.)	**Err.**	**Cohen's f^2**	β (sig.)	**Err.**	**Cohen's f^2**	β (sig.)	**Err.**	**Cohen's f^2**
Intercept	0.13 *	-0.06		-0.04	-0.07		0.12 ***	-0.04	
Edu[4Y&CColl]	0.13 ***	-0.04	0.03	0.05 *	-0.02	0.01	0.07 ***	-0.02	0.02
Edu[HS&SColl]	0	-0.04	0.03	0.01	-0.04	0.01	-0.03	-0.02	0.02
Gender[Male]	-0.02	-0.01	0.01	-0.04 *	-0.02	0.00	-0.03 ***	-0.01	0.01
Age[26-30]	-0.05	-0.04	0.00	-0.01	-0.03	0.01	-0.06 ***	-0.02	0.00
Pol[Indep.]	0.06 ***	-0.02	0.01	0.11 ***	-0.02	0.03	0.06 ***	-0.01	0.02
Pol[Other]	0.04	-0.02	0.01	0.15 ***	-0.03	0.03	0.08 ***	-0.01	0.02
Pol[Repub.]	0.08 ***	-0.02	0.01	0.13 ***	-0.04	0.03	0.10 ***	-0.01	0.02
N		2450			1297			3747	
R^2/Adj. R^2		0.16/0.15			0.14/0.13			0.15/0.14	

Table 2. OLS regression on error in credibility rating compared to science experts' average rating after recoding and non-significant rows omitted. The reference for education, gender, age and political leaning are: Graduate degree, Female, 18-25 and Democrat. Numbers in green are negative coefficients with significant p-values contributing to less error; numbers in red are vice-versa. Here, Cohen's f^2 and adjusted R^2 are the effect size of each variable and each model respectively. Conventionally, Cohen's f^2 of 0.02, 0.15, and 0.35 are termed small, medium, and large, respectively.

	Expert[Journalism]								
	Student			Upwork			Stud.+Upwork		
	β (sig.)	**Err.**	**Cohen's f^2**	β (sig.)	**Err.**	**Cohen's f^2**	β (sig.)	**Err.**	**Cohen's f^2**
Intercept	0.27 ***	-0.06		0.11	-0.07		0.26 ***	0.04	
Edu[4Y&CColl]	0.11 ***	-0.04	0.03	0.04 *	-0.02	0.00	0.06 ***	0.02	0.02
Edu[HS&SColl]	-0.03	-0.04	0.03	0.01	-0.04	0.00	-0.05 ***	0.02	0.02
Gender[Male]	-0.03 *	-0.01	0.01	-0.04 ***	-0.02	0.00	-0.03 ***	0.01	0.01
Age[26-30]	-0.06	-0.04	0.00	-0.03	-0.03	0.01	-0.06 ***	0.02	0.00
Pol[Indep.]	0.07 ***	-0.02	0.02	0.12 ***	-0.02	0.03	0.07 ***	0.01	0.02
Pol[Other]	0.06 **	-0.02	0.02	0.15 ***	-0.03	0.03	0.08 ***	0.01	0.02
Pol[Repub.]	0.10 ***	-0.02	0.02	0.14 ***	-0.04	0.03	0.11 ***	0.01	0.02
N		2450			1297			3747	
R^2/Adj. R^2		0.12/0.12			0.10/0.09			0.11/0.11	

Table 3. OLS regression on error in credibility rating compared to journalism experts' average rating after recoding and non-significant rows omitted. The reference for education, gender, age and political leaning are: Graduate degree, Female, 18-25 and Democrat. Numbers in green are negative coefficients with significant p-values contributing to less error; numbers in red are vice-versa. Here, Cohen's f^2 and adjusted R^2 are the effect size of each variable and each model respectively. Conventionally, Cohen's f^2 of 0.02, 0.15, and 0.35 are termed small, medium, and large, respectively.

College/Vocational Training" into another. We also divided raters into "18-25", "26-30", and "31+" age groups.

4.2.1 Democrats, males, ages 26–30, and people with higher education levels have greater alignment with experts on climate science. Among our variables, consistent across all models, crowd raters with a non-Democrat political leaning had higher error in their assessment (where error is alignment with the experts in the particular model). In addition, males had lower error compared to females; the difference is small but significant in all the models except one. Among age groups, people aged 26–30 had lower error compared to those aged 18–25; however those values are only significant in the omnibus models. Other age ranges had no significant results. On the other hand, crowd raters with a four-year college or community college degree had higher error compared to those with a graduate degree. Surprisingly, raters with a high school degree or some college experience had lower error compared to those with a graduate degree in one of our models (`Student+Upwork` compared with `Journalism`). This may be because the majority of our crowd raters in the `Student` group are assumed to still be in college, and perform relatively well due to exposure to journalism

	Count	Student	Upwork	Expert[sci]	Expert[jour]
Opinion	8	**0.398**	**0.477**	**0.742**	0.525
Analysis	8	0.355	0.440	0.625	**0.809**
News	32	0.311	0.339	0.518	0.537
Left	6	0.251	0.304	0.236	**0.597**
Center	24	0.095	0.141	**0.328**	0.136
Right	15	**0.330**	**0.322**	0.247	0.508

Table 4. IRR across article genres and political leaning of article sources. Here, the numbers in bold represent the highest IRR for each rater group across article genres/political leaning.

and media studies. Thus in addition to the aspects of potential bias due to political orientation, potentially exacerbated in the case of climate change news as we expected, we find that the issue of formal training and education is important to consider.

4.3 RQ3: Rating Performance According to Article Type

In this section, we investigate specifically how the genre of an article as well as the political leaning of the publication result in differences between expert and crowd ratings. Given the difficulty that Americans have with factual and opinion statements within news articles, we first consider article *genre*. As explained earlier, journalism experts additionally classified the genre of articles in our dataset, applying "Opinion", "Analysis", and "News". We used a majority vote by the journalists to categorize 48 out of 50 articles into their respective genres. Across News and Opinion, the journalism experts had an IRR of 0.97; but when adding Analysis as a third category, the IRR went down to 0.71.[5]

The second area of interest is the *political leaning* of the publication behind an article. Using Media Bias/Fact Check, a site that classifies media sources on a political bias spectrum and that has been used in prior research [8], we re-coded their 7 categories into three higher-level categories of left, center, and right resulting respectively in 6, 24, and 15 articles from our dataset (5 were omitted because they had no entry in Media Bias/Fact Check). From an article source perspective, articles from both right- and left-leaning sources have higher IRR from the crowd than those in the center (see Table 4). This suggests that annotators might have used the leaning of sources as shortcuts to identify credibility, given how political lean today equates with believing in or denying climate change [67].

We examined how our crowd groups evaluated credibility in relation to experts for the two sets of article types. Using our previous approach of correlation analysis, we looked at the correlation of credibility between pairs of groups. For this comparison, instead of varying the number of crowd raters from 1-25, we sampled 25 crowd raters 100 times and averaged the resulting correlations into a single metric. To combine p-values, we show the statistics as a percentage of times p was significant in the samplings [5]. The pairs of groups that shows significant correlations ($p<0.05$) in more than 70% of the samplings are the values of interest (see Figures 5 and 6), where 70% is a heuristic we chose to reduce clutter in the figures. However, we note that some aspects of the data have imbalance, particularly for the number of experts (n=3), the number of articles in the Opinion or Analysis genres (n=8), and from left-leaning sources (n=6). The following results need to be considered in light of these constraints.

[5] Separately, we wondered whether our crowd raters could label genre. When asked to consider just News versus Opinion, IRR was lower at 0.43 for Student and 0.49 for Upwork but the majority assessment of each crowd group was 100% aligned with experts. Most articles labeled "Analysis" by journalists were labeled "News" by the crowd groups.

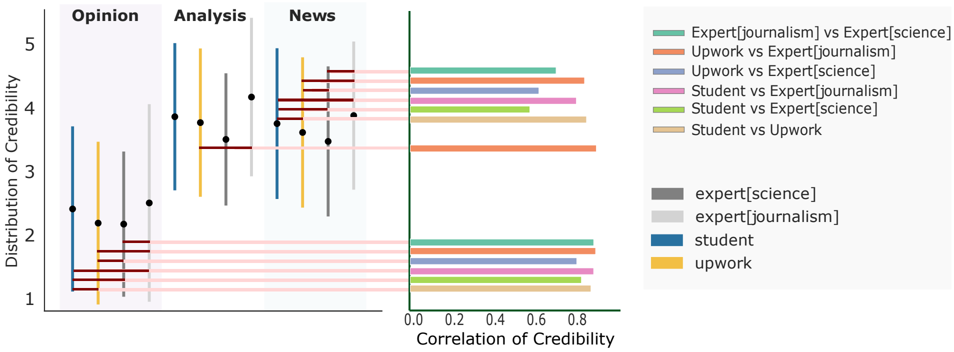


Fig. 5. This figure shows the average credibility rating and standard deviation for each of the four rater groups broken down by article genre of opinion/analysis/news (on the left side), along with correlation analysis results between pairs on the right side. The presence of a bar on the right means that a pair has significant (p<0.05) correlation of credibility in more than 70% of the crowd samplings. For the correlation analysis, we sampled crowd raters with n=25, sampling for 100 times, computing correlations each time and then averaging the correlations. Note that, number of articles for some categories are skewed (Opinion = 8, Analysis = 8, and News = 32).

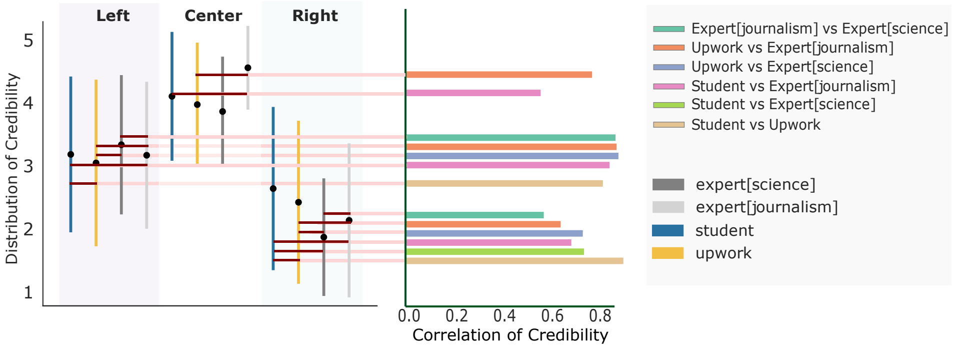


Fig. 6. This figure shows the average credibility rating and standard deviation for each of the four rater groups broken down by article source of left/center/right (on the left side), along with correlation analysis results between pairs (on the right side). The presence of a bar on the right means that a pair has significant (p<0.05) correlation of credibility in more than 70% of the crowd samplings. For the correlation analysis, we sampled crowd raters with n=25, sampling for 100 times, computing correlations each time and averaging the correlation. Note that, number of articles for some categories are skewed (Left = 6, Center = 24, and Right = 15).

4.3.1 Crowd groups correlated more with experts in the case of opinion articles. Among the different news genres, our tests suggest that crowd groups had higher correlation with both groups of experts on rating the credibility of Opinion articles. When it came to News articles, correlation of both crowds dropped with scientists but not with journalists. While scientists and journalists were somewhat correlated in the case of News, we saw differences between their average ratings, with journalists being more positive overall. This is even more pronounced in the case of Analysis, where journalists regarded these articles all relatively highly. In this genre, there was less correlation across rater groups. We explore potential reasons for this in RQ4.

4.3.2 Crowd groups correlated more with experts in articles from left-leaning publications. Along political lines, ratings of both crowd groups had higher correlation with the experts on articles from left publications. For articles from center publications, we only saw significant correlations between crowds and journalists; meanwhile, scientists and journalists disagreed. We also saw a high average rating from journalism experts for center publications, which may come from a professional experience and training that aligns more closely with center, non-partisan sources. This possibility is also explored in greater detail in RQ4. For right-leaning publications, both expert groups gave these articles low ratings on average, as expected, with science experts providing the lowest average rating. Interestingly, while crowd groups were highly correlated with each other, they had lower correlation with experts, and experts also had lower correlation between each other.

4.4 RQ4: Comparing Science and Journalism Experts

In order to understand why science versus journalism experts differ in their credibility assessments and how this might further illuminate crowd differences, we conducted a deep qualitative analysis of optional, open-ended explanations experts gave for their different credibility ratings. In total, the 3 scientists gave 82 explanations across the 50 articles, while the 3 journalists gave 65 explanations.

Initially, one of the authors conducted open coding across all of the explanations using a grounded theory method to develop an initial set of 38 codes of both negative and positive expert criteria [65]. All authors then discussed the codes while looking at examples of explanations, resulting in some codes being renamed and others being split apart or merged together. The authors also worked together to group the codes into high-level categories, some of which have a rough mapping onto existing principles of journalism [49]. After additional iterations of discussion and re-coding of the explanations, we arrived at the 8 high-level categories in Table 5. Each category is comprised of several lower-level criteria that are either positive or negative with regards to impact on credibility. For example, the code "accurate, based in facts[+]" under `Accuracy` means that an expert mentioned accuracy as a positive association to article credibility in their explanation.

4.4.1 Journalists primarily cite Publication Reputation while scientists consider multiple criteria. Overall, we found that experts mentioned `Accuracy` and `Publication Reputation` most frequently (48 times) closely followed by `Credible Evidence/Grounding` (45 times) and `Impartiality` (44 times). However, there were differences when we compared journalists versus scientists. By far the most cited criteria for journalists was `Publication Reputation` (Figure 7). We saw numerous cases where the journalists would either dismiss or trust the contents of an article based on the publication's brand and reputation: "*The Hill, while a crappy publication, has brand recognition that gives it more credibility. Without it the credibility ranking would be lower.*" Journalists were also more likely than scientists to mention criteria related to `Website Aesthetic` ("*serial killer font*") and `Professionalized Practices and Standards`, such as presence or lack of structured information such as a dateline and low writing/editing quality: "*...use of exclamation marks and bad writing overall reduced credibility in my mind.*"

In comparison, scientists were most likely to cite issues related to `Credible Evidence/Grounding`, such as the presence or lack of citations, quotes from experts, or other evidence: "*A partisan article...failing to include credible sources' comments on the decision.*" Scientists also mentioned `Impartiality` often, primarily to comment on neutrality of tone. Journalists mentioned impartiality frequently as well but were more likely to discuss it in terms of "both sides" coverage, in both a positive ("*Credibility enhanced by links to other publications and presentation of both sides*

Accuracy	**Impartiality**	**Completeness of Coverage**	**Originality and Insight**
accurate, based in facts[+] inaccurate representation of facts/scientific consensus[-] misleading images[-] misleading headline[-] sensationalist headline[-] hyperbolic language[-] cherrypicking/misleading[-]	neutral, nonpartisan tone/ lack of attacks or injected opinion[+] balanced/both sides of debate[+] goes against source/author's perceived bias/hurts their own cause[+] biased language, partisan, opinionated rant without substance[-] imbalanced/lack of both sides of debate[-] goes along with perceived bias[-]	provides context/explanation[+] thorough/in-depth as opposed to light coverage[+] Lack of context[-] light/cursory coverage[-]	provides insight/informed implications[+] lack of quality discussion/analysis/ insight[-] lack of original reporting[-] poor interpretation/ uninformed implications[-]
Credible Evidence/Grounding	**Publication Reputation**	**Professionalized Practices and Standards**	**Website Aesthetic**
references a credible source and/or confirmation by credible source [+] quotes from experts[+] cites credible scientific study[+] includes data/charts/image evidence [+] lack of citation[-] lack of quotes from experts[-] has citation but of bad science or low credibility study/source[-] facts refuted by credible source/ commonly known as debunked[-]	well-known/credible source[+] credible/expert author[+] low quality source[-] unknown/non-mainstream source/brand[-] biased source[-]	dateline clearly marked[+] clear article/source standards[+] clearly labeled as opinion when it is an opinion[+] authoritative, professional writing[+] lack of dateline[-] low writing/editing quality[-] personalization of language/ non-professional language[-]	poor font choice[-] bad page layout[-]

Table 5. Qualitative codes under 8 major categories. +/- symbols inside the brackets show their polarity on credibility. See Appendix C for example notes and their corresponding codes.

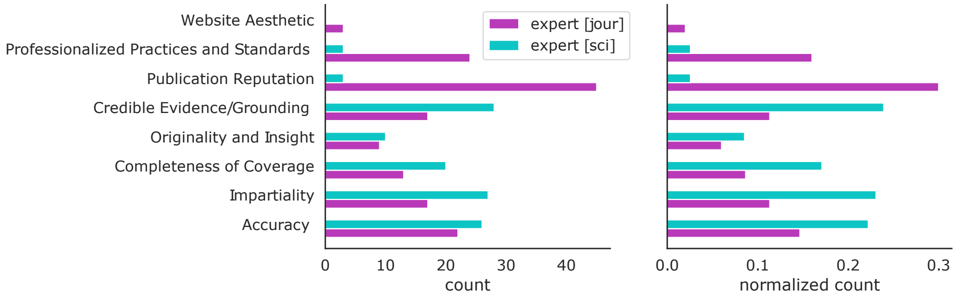


Fig. 7. Frequency of the categories in expert explanations for journalists versus scientists. On the left are raw counts and on the right, the counts are normalized by the number of explanations made by journalists versus scientists in total.

of argument/critics views...") or negative way (*"Links add to credibility. However, there is no opposing/contrarian voices in this story.")* Finally, scientists were also more likely to cite Accuracy and would sometimes rely on their personal knowledge about the science to evaluate the article: "*I study satellite imagery...A really poor study, repeatedly debunked.*"

We performed a series of regressions with experts' credibility rating as the outcome variable and their codes divided into positive and negative factors as independent variables. Table 6 shows the result of our model for the three combinations of science experts, journalism experts, and then the two sets of experts combined. We tested for multicollinearity in the data and found no evidence of it (Variation Inflation Factor < 1.2, ∀ factors). The beta scores with significance

	Science		Journalism		Sci + Jour	
	β (sig.)	**std. err.**	β (sig.)	**std. err.**	β (sig.)	**std. err.**
Intercept	3.54 ***	(0.15)	3.66 ***	(0.13)	3.62 ***	(0.09)
Completeness of Coverage[+]	0.75 *	(0.32)	0.38	(0.39)	0.57 *	(0.25)
Credible Evidence/Grounding[+]	0.57 *	(0.28)	0.13	(0.42)	0.40	(0.23)
Publication Reputation[+]	0.18	(0.67)	0.78 *	(0.34)	0.59 *	(0.26)
Accuracy[-]	-0.74 ***	(0.20)	-0.71	(0.57)	-0.78 ***	(0.20)
Impartiality[-]	-0.81 ***	(0.23)	-0.71	(0.51)	-0.82 ***	(0.22)
Originality and Insight[-]	-1.01 *	(0.40)	-0.53	(0.54)	-0.74 *	(0.32)
Credible Evidence/Grounding[-]	-0.99 ***	(0.26)	-0.71	(0.87)	-0.94 ***	(0.26)
Professionalized Practices and Standards[-]	-0.96	(0.55)	-0.89	(0.49)	-0.81 *	(0.33)
R^2	0.55		0.30		0.39	

Table 6. Regression on credibility using qualitative codes. Non-significant rows have been omitted.

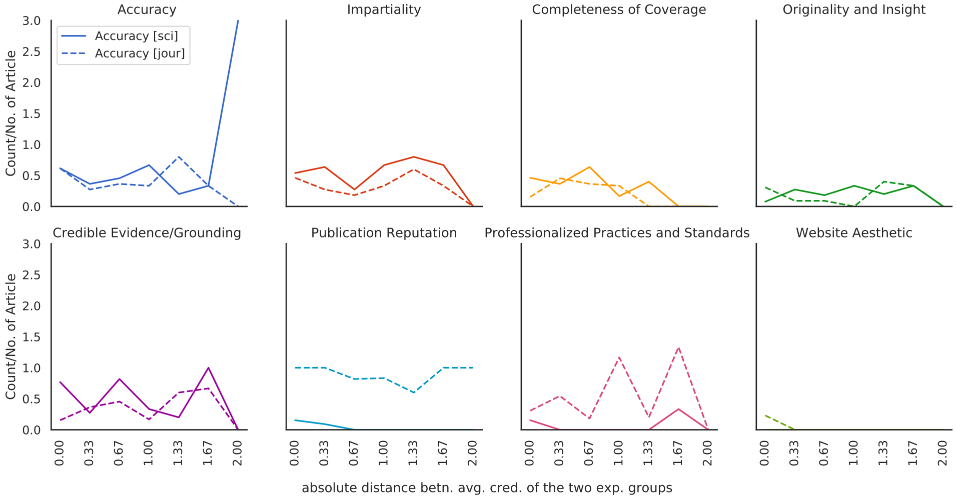


Fig. 8. Count of occurences of the codes normalized by the number of articles for articles with differing absolute distance between science and journalism [abs(avg(sci) - avg(jour))] experts' average credibility ratings.

demonstrate that science experts cite multiple categories whereas journalists tend to focus on Publication Reputation, primarily as positive evidence. However, for science experts, among the different criteria that could increase or decrease their credibility perception, only Credible Evidence/Grounding had both significant positive and negative impact. The remaining categories only boosted their perception of credibility (Completeness of Coverage) or only negatively influenced it (Accuracy, Impartiality, Originality and Insight).

4.4.2 Major disagreements arise due to emphasis on Accuracy versus Publication Reputation. In Figure 8, we show how often a particular criteria is provided by scientists versus journalists as the absolute difference between their average ratings for an article increases. We can see that as disagreements between scientists and journalists grow, their rationales diverge, with scientists citing Accuracy more, and journalists citing Publication Reputation and Professionalized Practices and Standards more.

We inspected some examples of articles with high absolute differences in ratings between the experts to illustrate how these differences emerged. For instance, in one case, scientists rated an article by the Daily Wire, an outlet considered to have a "right bias with mixed factual reporting" according to the site Media Bias/Fact Check, as considerably more credible than journalists did

(1.67 difference in average ratings). The article was reporting on an academic publication, leading one scientist to write "*reasonable reporting on a study that has some issues with reaching claims*" and to give it a 3 out of 5. Journalists were considerably more harsh, taking the article to task for issues such as lack of `Originality and Insight` and `Professionalized Practices and Standards`: "*...it's a news story that cites a study but has no real original or live onsite reporting. Lack of deadline undermines credibility.*" They also mentioned `Publication Reputation`, with one person stating the article's credibility was "*undermined by association [with] previous content deemed not credible*" on the site.

In another case, we saw journalists this time giving an article by BBC News a higher rating (5 out of 5 across the board), while scientists all gave the article a 3 out of 5. Unsurprisingly, journalists mentioned `Publication Reputation`, with one person saying that credibility was "*...enhanced by association with BBC brand.*" However, scientists found issues with `Accuracy`, calling out the piece for misleading images and a misleading headline: "*the title including the word 'hothouse' can be misleading as it suggests a runaway global warming, which is not possible on Earth.*"

These examples point to the shortcuts that journalists sometimes employ by focusing on an article's publication or more superficial elements of style and presentation, as opposed to the contents of an article. This may be necessary in cases when they cannot easily consult the underlying scientific source and do not have access to the deep domain knowledge that scientists can draw upon. This may be why we saw journalists giving uniformly high ratings to center-leaning publications in RQ3. This may also explain why crowd raters tended to agree with journalists more.

Finally, we noticed a few major differences in ratings stemmed from differences in interpreting genres of news articles. In several instances, we saw scientists giving lower scores to articles that would be considered "straight news", or news articles that concisely and impartially report facts about an event, while journalists gave them a 5. For example, in one article labeled News by the journalists where there was a difference of 1.3 between scientist and journalist ratings, a scientist gave the following rationale for their rating of 3: "*Neutral account of incident, no insight provided.*" This may be why we see scientists invoking `Completeness of Coverage` at a higher rate than journalists, as journalists may perceive a concise article without in-depth coverage as a valid piece of journalism. This could also explain why journalists overall gave high ratings for the genres of analysis and news in RQ3 compared to scientists.

5 DISCUSSION

This study investigated several sources of difference between the layperson assessment of news credibility and that of experts in science and journalism, all towards the goal of informing crowdsourced processes for news credibility assessment at scale. RQ1 affirmed that crowds do not always agree with experts, and experts do not always agree amongst themselves. If the goal is to align crowds to experts, we find that it takes about 15 crowd raters to achieve high correlation, after which correlation begins to plateau. However, this number might be reduced if we tailor crowds and tasks, given our findings in RQ2 and RQ3.

Interestingly, we find that the `Upwork` crowd has a slightly higher correlation with experts than the `Student` group, many of whom presumably took coursework in media literacy or journalism. However, while `Upworkers` have a more varied demography than our `Student` group, they also likely have high rates of digital literacy as online freelancers [48]. When we examine demographics more carefully in RQ2, we find that Democrats, males, ages 26–30, and people with higher education levels across both crowd groups have greater alignment with experts. However, some results are likely specific to the topic, given that the Republican platform currently questions climate change. Other factors such as gender may be ones in which it may not be desirable to have biased representation.

Delving into article types was the focus of RQ3, which laid some groundwork for task suitability. When it comes to *genre*, both groups of crowd raters were more correlated with experts on opinion articles. Along political lines, crowd groups were more correlated with experts on articles from left-leaning sources. These results suggest that the crowd may have the ability to replace experts' annotations in certain article types but not others. In addition, it may be that some difficulties for raters arise from the lack of visual cues such as genre labeling in U.S. mainstream media [33]. Without being labeled or well understood, readers might need to rely on structural aspects such as article genre classification when the style is difficult to interpret, and experts themselves cannot always agree. Finally, given our findings in RQ4, some news articles that conduct original research or report on new scientific findings might require subject matter experts who can assess accuracy.

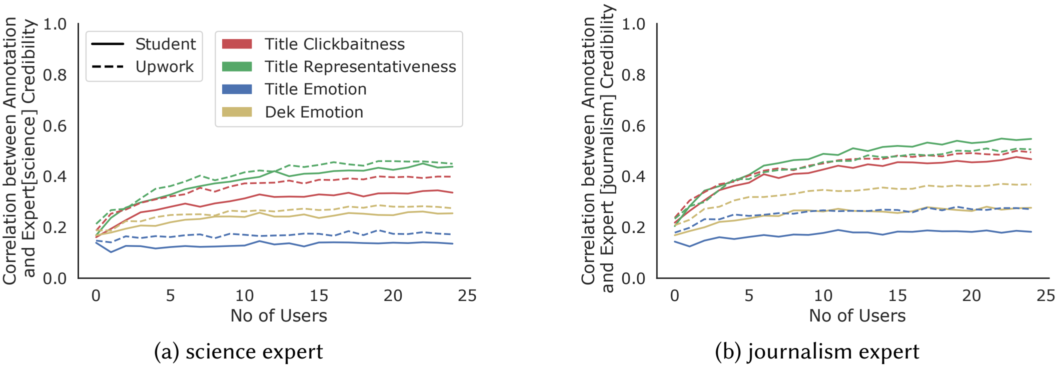
5.1 Tailoring Tasks to Align on Credibility Criteria

While we cannot expect crowds to always be capable of evaluating `Accuracy`, results from RQ4 pointed to more attainable ways to evaluate news articles that experts also use, namely the inclusion of `Credible Evidence/Grounding` used by scientists and `Publication Reputation` used by journalists. Though over-emphasis on `Publication Reputation` by journalism experts may seem to be a red flag, it is a way for the non-experts of a domain such as climate science to initiate their investigation of credibility, much as scientists have preconceptions about the work from certain scholarly journals over others.

Given our findings that domain experts use different criteria to judge credibility and that these differences may surface among crowds, a future line of work could seek to reduce crowd disagreement both within itself and with certain experts by aligning to a particular set of domain-relevant criteria. For example, one might ask the crowd to label specific components of an article that may *signal* credibility, rather than broadly asking about credibility itself. This forces raters to focus on aspects such as `Publication Reputation` or `Credible Evidence/Grounding` that align with expert assessments as opposed to allowing raters to reduce the broad credibility question into a scale according to a dimension of their own choosing or instinct.

Indeed, prior research has shown that crowds perform well at assessing publication reputation [53], and there exists a wide set of such source and message characteristics or *signals* of potential trust from a reader's perspective, ranging from organization standards to the reader's capacity for engagement [13, 17, 22, 42, 55, 56]. Other work has examined features directly in the article that may signal credibility—including title structure and proper noun [31], article content (e.g., emotional tone) or context (e.g., citation to reputable sources) [74]—as well as secondary characteristics (e.g., source attractiveness [50]). Even in the news credibility context, research indicates that crowd and journalists' evaluation of information accuracy differ in their incorporation of signals [10]. If a complex construct like credibility can be distilled into a cluster of simpler signals, such as the perception of emotion in an article's title, and further designed to be in alignment with expert judgments, annotators may prove to be far more reliable in the completion of those tasks rather than more complicated assessments. This sub-task strategy is familiar in complex crowdsourcing systems [37].

To pilot this concept, we applied this reasoning to our own study, which asked the crowd groups to additionally label four credibility signals that had previously been identified as potential indicators of expert credibility [74] on all 50 articles. Again, because our scope was for large-scale credibility assessment, we also looked for quickly answerable questions. The questionnaire included three signals related to the title of the article: the degree to which it is "clickbait", its level of emotion, and the representativeness of the title in comparison to the rest of the article. The fourth asked about



(a) science expert

(b) journalism expert

Fig. 9. Correlation between article signals and 2 expert groups' credibility rating.

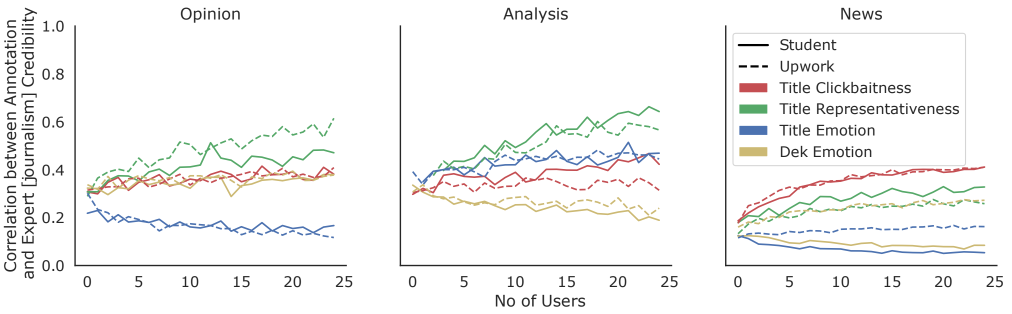


Fig. 10. Correlation between article signals and journalism expert credibility rating for opinion/analysis/news.

the level of emotion in the "dek", or short summary of the article beneath the title[6]. Seen through the work of RQ4, the emotion signals relate to our expert category of `Impartiality` while our representativeness of title and clickbait title correspond to our expert categories of `Accuracy` and `Professionalized Practices and Standards`.

We report that the signals we tested overall resulted in poor to moderate correlations with expert credibility assessments. We found that crowd groups' title representativeness scores had the highest alignment with experts' credibility ratings (see Figure 9(a) and 9(b)), yet had the lowest IRR between raters (0.16 for `Student` and 0.19 for `Upwork`). In contrast, both the emotion questions had low correlation with experts' credibility ratings but had the highest IRR (e.g., for title emotion, IRR was 0.47 for `Student` and 0.52 for `Upwork`). We also saw that overall, Upwork workers have higher correlation to both groups of experts in comparison to `Students`, except for title representativeness against the journalism experts.

But now, taking our qualitative categories and subcategories based upon expert rationales, we might frame the question differently, seeking "misleading headline" or "sensationalist headline" under the `Accuracy` category in addition to preferencing other categories altogether. Another strategy may be to combine our insights regarding article type along with signals for credibility. Figure 10 shows the correlation between journalists' credibility rating and crowd ratings on credibility signals broken down across three article genres. Crowd workers have higher correlation with journalism experts on all our signals for *Opinion* and *Analysis* articles in contrast to *News*. This

[6] See Appendix D for how we defined these terms for the crowds.

pattern suggests some of the signals can be more useful in particular cases (e.g., title representativeness has a correlation as high as 0.6 for *Opinion* and *Analysis* articles, while title clickbaitness is the most correlated signal for the *News* genre). More work is needed to find credibility signals that align well with expert criteria and that are also stable among some subset of the crowd before this approach can be practically used in production crowdsourced processes.

5.2 Design Implications

Our current work has implications for designing processes for crowdsourcing news credibility. We summarize them below.

5.2.1 Recruitment and Training of the Crowd. Designers have the opportunity to control the participants involved in crowdsourcing at two levels: demographic filtering during recruitment and training for secondary improvement. In other words, our results imply that a combination of *person-oriented* strategies (e.g., filtering by demographics), followed by *process-centric* strategies (e.g., training raters by emphasizing what signals they should consider) can facilitate high-quality, at-scale credibility assessment. These results are in line with prior work pointing at the advantages of person- and process-centric strategies for crowd-sourcing qualitative coding [47], which includes tasks that are often quite subjective in nature and, thus, prone to conflicting interpretations. Our approach to credibility of news articles is indeed a blend of subjective and objective assessments.

Based on the performance of the two crowd groups with differing demographic backgrounds, our findings also suggest that 15 ratings provide enough stability in the result. However, the difference in errors based on background suggests that recruiters can employ certain *person-centric* filtering mechanisms to enforce specific criteria in their systems. For example, filtering out certain education levels may serve some purpose for the system designers. At the same time, designers should be aware of how such a filtering mechanism may bias the system. Additionally, criteria used by the expert groups (demonstrated in our RQ4 results), could serve as training for the crowd, offering a host of process-oriented tactics for designers to employ on their crowd-rater workforce. For example, questionnaires can be devised to identify a baseline of crowd raters' expertise in credibility evaluation. Based on the expertise, different training mechanisms can be targeted towards each group to improve their deficiency (e.g., literacy programs to improve accuracy or impartiality identification). This training should not be limited towards understanding only the principles of journalism; rather it could show how subject matter experts identify and distill reliable evidence.

5.2.2 Considerations for Comparison with Experts. Given the differences in evaluation criteria and corresponding credibility ratings, designers have to consider which group of experts they want to emulate in the system. This consideration is in effect throughout the design process. For instance, to emulate behavior close to the journalism experts, system developers may employ specific strategies in their recruitment process. However, desirable expertise can vary case-by-case. For example, with science news, it might be desirable to have crowd ratings closer to a science expert's understanding of the subject matter (`Credible Evidence/Grounding`) while for breaking information news consumers may appreciate ratings that reflect journalistic expertise in verifying source quality (`Publication Reputation`). A greater understanding of desirable expertise in different news stories would further help future design.

5.2.3 Task Suitability. Analysis of the article types suggests that some articles may require subject matter expertise while others may be reliably assessed by the crowd. For article types where crowds have high disagreement with experts, alternate approaches can be devised including training focused on particular flaws or a very tailored set of questions. Our study focused on a topic area in which U.S. Democrats have been shown to have a stronger relationship to credible information. A closer examination of other topics, with different kinds of polarization and expertise, is needed. Vaccine hesitancy, for instance, is an issue with traction across the political spectrum as are a number of conspiracy theories, with the former attitude now represented in general crowds in contrast to the continued fringe nature of the latter [34, 70]. The relationship between credibility and a strong partisan perspective may not exist in these cases, but we may find other factors of belief such as extremism, in addition to relevant expert criteria that can frame tasks better. Another consideration in task suitability is the task difficulty where even expert groups diverge. In these cases, policy decisions may need to be made regarding which expertise is more relevant before designing crowd tasks.

Overall, a successful crowdsourcing approach requires that tasks to be designed carefully with specific crowds, content, and experts all in mind. In section 5.1 we propose an approach that focuses on signals as an example. We note that this approach requires us to find signals that not only align with expert judgments but also are possible for crowds to locate and assess in a reliable and consistent manner.

5.3 Limitations

Our analysis suffers from several limitations. First, the result is limited by our dataset derived from only a popular set of articles on a particular topic of science news, an area in which domain experts largely agree. Making general claims from such limited data would be inaccurate as we have explained throughout, but we can infer that the corresponding relationship among experts and raters might be at least as complicated as we found, if not more so. Second, annotation from our recruited populations of both crowd and expert groups are also limited by quantity and demographic/expertise background. In particular, differences between our two expert groups could have been due to our restricted sample. Third, some of the expert results are drawn from an incomplete set of expert criteria, and our codebook for expert criteria is constructed from the authors' interpretation; thus, conclusions there also have their limitations.

6 CONCLUSION

In this work, using the domain of climate news, we dive into the notion of crowdsourcing credibility through a series of analyses on its main components: the makeup of the *crowd*, the scope of *tasks* that the crowd is assigned, and the subject area *expert criteria* in question. In particular, we explore characteristics of the "crowd," in terms of traits such as background, demographics, and political leaning, and whether they have bearings on task performance. We show this in a comparison between ratings made by students and others recruited through journalism networks versus crowd workers on UpWork. We also interrogate the nature of the crowdsourcing task itself, finding that the genre of the article and partisanship of the publication has different relationship to both crowds and experts. This led us to better understand the reasoning of experts themselves. In our case, we looked at how experts in journalism versus experts in science have different ways to assess article credibility based on the factors such as `Credible Evidence/Grounding` and `Publication Reputation`. Disagreement among raters is neither always bad nor always about their capacities, but at times about suitability of the task [2] and about the particular subject area expertise in question as well. By investigating the variability introduced by all these components, we point towards how the design of crowd assessments to approximate expert-level credibility can be made more robust.

7 ACKNOWLEDGEMENTS

This paper would not be possible without the valuable support of the Credibility Coalition, with special thanks to Caio Almeida, An Xiao Mina, Jennifer 8. Lee, Rick Weiss, Kara Laney, and especially

Dwight Knell. Bhuiyan and Mitra were partly supported through National Science Foundation grant #IIS-1755547.

REFERENCES

[1] William RL Anderegg, James W Prall, Jacob Harold, and Stephen H Schneider. 2010. Expert credibility in climate change. *Proceedings of the National Academy of Sciences* 107, 27 (2010), 12107–12109.

[2] Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (2015), 15–24.

[3] Mahmoudreza Babaei, Abhijnan Chakraborty, Juhi Kulshrestha, Elissa M Redmiles, Meeyoung Cha, and Krishna P Gummadi. 2019. Analyzing Biases in Perception of Truth in News Stories and Their Implications for Fact Checking.. In *FAT*. 139.

[4] Mevan Babakar. 2018. Crowdsourced Factchecking.

[5] Betsy Jane Becker. 1994. Combining significance levels. *The handbook of research synthesis* (1994), 215–230.

[6] Joshua Becker, Ethan Porter, and Damon Centola. 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences of the United States of America* 166, 22 (2019), 10717–10722. https://doi.org/10.1073/pnas.1817195116

[7] Brooke Borel. 2015. The problem with science journalism: we've forgotten that reality matters most. *The Guardian* (Dec 2015). https://www.theguardian.com/media/2015/dec/30/problem-with-science-journalism-2015-reality-kevin-folta

[8] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.

[9] Mohamad Adam Bujang and Nurakmal Baharum. 2016. Sample size guideline for correlation analysis. *World* 3, 1 (2016).

[10] Cody Buntain and Jennifer Golbeck. 2017. Automatically Identifying Fake News in Popular Twitter Threads. *Proceedings - 2nd IEEE International Conference on Smart Cloud, SmartCloud 2017* (2017), 208–215. https://doi.org/10.1109/SmartCloud.2017.40 arXiv:1705.01613

[11] Davide Ceolin. 2019. Conference Presentation: On the Quality of Crowdsourced Information Quality Assessments. https://drive.google.com/a/hackshackers.com/file/d/1AJmFmRqEhdhSIZLwhXT_1bzStXfV-hVf/view?usp=drive_open&usp=embed_facebook

[12] Steven H Chaffee. 1982. Mass media and interpersonal channels: Competitive, convergent, or complementary. *Inter/media: Interpersonal communication in a media world* 57 (1982), 77.

[13] Shelly Chaiken. 1987. The heuristic model of persuasion. In *Social influence: the ontario symposium*, Vol. 5. Hillsdale, NJ: Lawrence Erlbaum, 3–39.

[14] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems* 50, 1 (2000), 1–18.

[15] Jaap J Dijkstra, Wim BG Liebrand, and Ellen Timminga. 1998. Persuasiveness of expert systems. *Behaviour & Information Technology* 17, 3 (1998), 155–163.

[16] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–11. https://doi.org/10.1145/3313831.3376232

[17] Jonathan St BT Evans. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* 59 (2008), 255–278.

[18] Facebook. 2020. Fact-Checking on Facebook: What Publishers Should Know. https://www.facebook.com/help/publisher/182222309230722. (Accessed on 01/14/2020).

[19] FactCheckEU. [n.d.]. FactCheckEU - 19 European media outlets are fact-checking the May 2019 European elections. https://www.factcheckeu.info/en/. (Accessed on 05/14/2020).

[20] Andrew J Flanagin and Miriam J Metzger. 2008. Digital media and youth: Unparalleled opportunity and unprecedented responsibility. *Digital media, youth, and credibility* (2008), 5–27.

[21] Fabrice Florin. 2010. Crowdsourced Fact-Checking? What We Learned from Truthsquad. *Mediashift* (2010).

[22] Brian J Fogg. 2003. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*. Citeseer, 722–723.

[23] Brian J Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 80–87.

[24] American Press Institute & The AP-NORC Center for Public Affairs Research. 2018. Americans and the news media: What they do–and don't–understand about each other. *The Media Insight Project* (2018).

[25] Cary Funk, Meg Hefferon, Brian Kennedy, and Courtney Johnson. 2019. Trust and Mistrust in Americans' Views of Scientific Experts. *Pew Research Center. https://www. pewresearch. org/science/2019/08/02/trust-and-mistrust-inamericans-views-of-scientific-experts* (2019).

[26] Cecilie Gaziano and Kristin McGrath. 1986. Measuring the concept of credibility. *Journalism quarterly* 63, 3 (1986), 451–462.

[27] Emma Grillo. 2020. What Does a Sports Desk Do When Sports Are on Hold? *The New York Times* (Apr 2020). https://www.nytimes.com/2020/04/05/reader-center/coronavirus-sports-reporting.html

[28] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (Jan 2019), 374–378.

[29] Naeemul Hassan, Mohammad Yousuf, Mahfuzul Haque, Javier A Suarez Rivas, and Md Khadimul Islam. 2017. Towards A Sustainable Model for Fact-checking Platforms: Examining the Roles of Automation, Crowds and Professionals. https://doi.org/10.1145/3308560.3316734

[30] Brian Hilligoss and Soo Young Rieh. 2008. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management* 44, 4 (2008), 1467–1484.

[31] Benjamin D. Horne and Sibel Adali. 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. (2017), 759–766. arXiv:1703.09398 http://arxiv.org/abs/1703.09398

[32] Carl Iver Hovland, Irving Lester Janis, and Harold H Kelley. 1953. Communication and persuasion. (1953).

[33] Rebecca Iannucci and Bill Adair. 2017. Reporters' Lab Study Results: Effective News Labeling and Media Literacy.

[34] Jonathan Kennedy. 2019. Populist politics and vaccine hesitancy in Western Europe: an analysis of national-level data. *European Journal of Public Health* 29, 3 (Jun 2019), 512–516. https://doi.org/10.1093/eurpub/ckz004

[35] Gary King and Richard Nielsen. 2019. Why propensity scores should not be used for matching. *Political Analysis* 27, 4 (2019), 435–454.

[36] Spiro Kiousis. 2001. Public trust or mistrust? Perceptions of media credibility in the information age. *Mass communication & society* 4, 4 (2001), 381–403.

[37] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.

[38] Michael Lucibella. 2009. Science Journalism Faces Perilous Times. *American Physical Society (APS) News* 18, 4 (Apr 2009). http://www.aps.org/publications/apsnews/200904/journalism.cfm

[39] Albert Mannes, Jack Soll, and Richard Larrick. 2014. The Wisdom of Select Crowds. *Journal of personality and social psychology* (2014).

[40] Albert E. Mannes, Richard P. Larrick, and Jack B. Soll. 2012. The social psychology of the wisdom of crowds.

[41] Aaron M. McCright, Katherine Dentzman, Meghan Charters, and Thomas Dietz. 2013. The influence of political ideology on trust in science. *Environmental Research Letters* 8, 4 (Nov 2013), 044029.

[42] Miriam J Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 2078–2091.

[43] Miriam J Metzger, Ethan H Hartsell, and Andrew J Flanagin. 2015. Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research* (2015), 0093650215613136.

[44] Philip Meyer. 1988. Defining and measuring credibility of newspapers: Developing an index. *Journalism quarterly* 65, 3 (1988), 567–574.

[45] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. 2018. *Can Americans Tell Factual From Opinion Statements in the News?*

[46] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations. In *Proc. ICWSM'15*.

[47] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proc. CHI'15*. ACM, 1345–1354.

[48] Kevin Munger, Mario Luca, Jonathan Nagler, and Joshua Tucker. 2019. Age matters: Sampling strategies for studying digital media effects.

[49] American Society of Newspaper Editors. 1975. ASNE Statement of Principles. https://members.newsleaders.org/content.asp?pl=24&sl=171&contentid=171. (Accessed on 01/14/2020).

[50] Daniel J O'Keefe. 2008. Persuasion. *The International Encyclopedia of Communication* (2008).

[51] Sheila O'Riordan, Gaye Kiely, Bill Emerson, and Joseph Feller. 2019. Do you have a source for that? Understanding the Challenges of Collaborative Evidence-based Journalism. In *Proceedings of the 15th International Symposium on Open Collaboration*. 1–10.

[52] Gordon Pennycook, Tyrone Cannon, and David G. Rand. 2018. *Prior Exposure Increases Perceived Accuracy of Fake News*. Number ID 2958246. https://papers.ssrn.com/abstract=2958246

[53] 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 (Feb 2019), 2521–2526.

[54] Gordon Pennycook and David G. Rand. 2019. *Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking.* Number ID 3023545. https://papers.ssrn.com/abstract=3023545

[55] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

[56] The Trust Project. 2017. Collaborator Materials. https://thetrustproject.org/collaborator-materials/. (Accessed on 01/14/2020).

[57] Soo Young Rieh and David R Danielson. 2007. Credibility: A multidisciplinary framework. *Annual review of information science and technology* 41, 1 (2007), 307–364.

[58] Robert M. Ross, David G. Rand, and Gordon Pennycook. 2019. Beyond "fake news": The role of analytic thinking in the detection of inaccuracy and partisan bias in news headlines. (2019), 1–22.

[59] Linda Schamber. 1991. Users' Criteria for Evaluation in a Multimedia Environment.. In *Proceedings of the ASIS Annual Meeting*, Vol. 28. ERIC, 126–33.

[60] Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7662–7669.

[61] Tracy Jia Shen, Robert Cowell, Aditi Gupta, Thai Le, Amulya Yadav, and Dongwon Lee. 2019. How Gullible Are You?: Predicting Susceptibility to Fake News. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. ACM, 287–288. https://doi.org/10.1145/3292522.3326055 event-place: Boston, Massachusetts, USA.

[62] Art Silverblatt, Donald C. Miller, Julie Smith, and Nikole Brown. 2014. *Media Literacy: Keys to Interpreting Media Messages, 4th Edition: Keys to Interpreting Media Messages.* ABC-CLIO.

[63] Henry Silverman. 2019. Helping Fact-Checkers Identify False Claims Faster - About Facebook. https://about.fb.com/news/2019/12/helping-fact-checkers/. (Accessed on 01/10/2020).

[64] Julianne Stanford, Ellen R Tauber, BJ Fogg, and Leslie Marable. 2002. *Experts vs. online consumers: A comparative credibility study of health and finance Web sites.* Consumer Web Watch.

[65] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. *Handbook of qualitative research* 17 (1994), 273–85.

[66] S Shyam Sundar. 1999. Exploring receivers' criteria for perception of print and online news. *Journalism & Mass Communication Quarterly* 76, 2 (1999), 373–386.

[67] S Shyam Sundar. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility* 73100 (2008).

[68] Cass R. Sunstein. 2006. When Crowds Aren't Wise. *Harvard Business Review* (Sep 2006). https://hbr.org/2006/09/when-crowds-arent-wise

[69] James Surowiecki. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Doubleday.

[70] Jan-Willem van Prooijen, André P. M. Krouwel, and Thomas V. Pollet. 2015. Political Extremism Predicts Belief in Conspiracy Theories. *Social Psychological and Personality Science* 6, 5 (Jul 2015), 570–578. https://doi.org/10.1177/1948550614567356

[71] Christian Wagner and Ayoung Suh. 2014. The Wisdom of Crowds: Impact of Collective Size and Expertise Transfer on Collective Performance. (Jan 2014), 594–603. https://doi.org/10.1109/HICSS.2014.80

[72] Lorraine Whitmarsh. 2011. Scepticism and uncertainty about climate change: Dimensions, determinants and change over time. *Global Environmental Change* 21, 2 (May 2011), 690–700.

[73] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (Oct 2010), 686–688. https://doi.org/10.1126/science.1193147

[74] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of The Web Conference 2018*. 603–612.

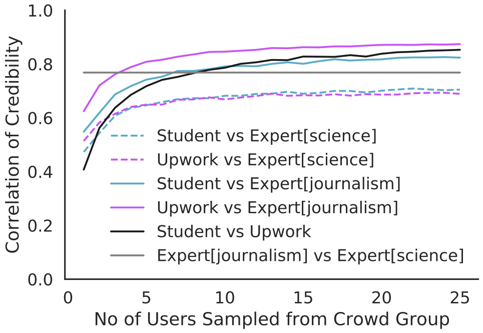


Fig. 11. Correlation of credibility ratings on the matched data. The lines show correlation among all pairs in four groups: 2 crowd and 2 expert groups. In each crowd group, we sample the number of raters from 1–25. For expert groups, we take all 3 ratings. Then we compute the Spearman ρ between the mean responses from each group on all 50 articles. The plot shows average ρ after 100 resamplings.

A RQ1: COMPARING UPWORK AND STUDENT CROWD RATERS TO EXPERTS WHEN CROWD GROUPS ARE CONTROLLED ON DEMOGRAPHY

The difference between the two crowd groups in RQ1 analysis could have been a result of some underlying issues including demographic variances. To account for this issue, we also performed RQ1 analysis by controlling for four demographic factors including Gender, Age, Education and Political Alignment. For this purpose, we created matching participants between Upwork workers and Students using *Match* function from ***R*** package *Matching* [7]. Because we had smaller number of users in the Upwork group, we matched them against Students with ties handled randomly. Due to duplicates, this method resulted in 21 unique students retained out of 49. We utilized `Mahalnobis Distance` as the matching criteria instead of `Propensity Score` because a recent work suggests that `Propensity Score` matching increases imbalance rather than decreasing [14, 35]. Figure 11 shows the correlation between and within crowd and expert groups on the matched data.

B NEWS ARTICLE DISTRIBUTION

Website	#	Website	#	Website	#
www.nytimes.com	5	www.bostonglobe.com	1	www.usatoday.com	1
www.breitbart.com	4	e360.yale.edu	1	www.dailykos.com	1
www.dailywire.com	4	www.economist.com	1	www.newsweek.com	1
www.theguardian.com	4	www.cnn.com	1	deadstate.org	1
www.npr.org	3	politi.co	1	expand-your-consciousness.com	1
www.foxnews.com	3	www.bbc.com	1	www.wsj.com	1
www.washingtonpost.com	2	arstechnica.com	1	www.iflscience.com	1
www.westernjournal.com	2	blogs.scientificAmerican.com	1	thehill.com	1
www.huffingtonpost.com	2	dailycaller.com	1	www.independent.co.uk	1
joeforamerica.com	1	www.smh.com.au	1	www.cbsnews.com	1

Table 7. Article distribution from the sources.

[7] https://sekhon.berkeley.edu/matching/Match.html

C SAMPLE OF EXPERT NOTES & QUALITATIVE CODES

Note	Codes
“A neutral discussion about the fight between left and right wing partisan on US President (lack of) role in the hurricane Florence disaster.”	(Impartiality) neutral, nonpartisan tone/lack of attacks or injected opinion[+]
“This story fails to include comments from independent scientists in the field or to provide necessary context for readers. For example, the study fails to account for more recent volcanic activity, and does not support its conclusion that climate models are overly sensitive to CO2. In addition, the story’s headline emphasizes that the study shows “no acceleration in global warming for 23 years” and this is presented as a challenge to model simulations. This is misleading, as no acceleration of the warming rate is expected to be seen in such a short timeframe. https://climatefeedback.org/evaluation/daily-caller-uncritically-reports-misleading-satellite-temperature-study-michael-bastasch/”	(Credible Evidence/Grounding) lack of quotes from experts[-] (Accuracy) misleading headline[-] (Originality and Insight) poor interpretation/uninformed implications[-] (Completeness of Coverage) lack of context[-] (Completeness of Coverage) light/cursory coverage[-]

Table 8. Sample notes from the experts and corresponding codes (high-level categories inside brackets). The colors show correspondence between Note and Codes.

D DEFINING SIGNALS

Clickbait. We provided following categories as examples of clickbait.

- Listicle ("6 Tips on ...")
- Cliffhanger to a story ("You Won’t Believe What Happens Next")
- Provoking emotions, such as shock or surprise ("...Shocking Result", "...Leave You in Tears")
- Hidden secret or trick ("Fitness Companies Hate Him...", "Experts are Dying to Know Their Secret")
- Challenges to the ego ("Only People with IQ Above 160 Can Solve This")
- Defying convention ("Think Orange Juice is Good for you? Think Again!", "Here are 5 Foods You Never Thought Would Kill You")
- Inducing fear ("Is Your Boyfriend Cheating on You?")

Representativeness. We suggested following categories on how an article can be unrepresentative.

- Title is on a different topic than the body
- Title emphasizes different information than the body
- Title carries little information about the body
- Title takes a different stance than the body
- Title overstates claims or conclusions in the body
- Title understates claims or conclusions in the body

Received January 2020; revised June 2020; accepted July 2020