INFORMED CROWDS CAN EFFECTIVELY IDENTIFY MISINFORMATION

February 10, 2022

ABSTRACT

Can crowd workers be trusted to judge whether news-like articles circulating on the Internet are wildly misleading, or does partisanship and inexperience get in the way? We assembled pools of both liberal and conservative crowd raters and tested three ways of asking them to make judgments about 374 articles. In a no research condition, they were just asked to view the article and then render a judgment. In an individual research condition, they were also asked to search for corroborating evidence and provide a link to the best evidence they found. In a collective research condition, they were not asked to search, but instead to look at links collected from workers in the individual research condition. The individual research condition reduced the partisanship of judgments. Moreover, the judgments of a panel of sixteen or more crowd workers were better than that of a panel of three expert journalists, as measured by alignment with a held out journalist’s ratings. Without research, the crowd judgments were better than those of a single journalist, but not as good as the average of two journalists.

Keywords  misinformation, social media, social platforms, crowdsourcing, survey equivalence

The rise of misinformation on social platforms is a serious challenge. About a fifth of Americans now obtain their news primarily from social platforms [Mitchell et al., 2020]; however, because anyone can post anything, a user can easily encounter misinformation [Vosoughi et al., 2018], [Grinberg et al., 2019]. Platforms have recently committed to countering the spread of misinformation, via methods such as informing news consumers that an item may be false or misleading, algorithmically demoting it, or removing it entirely [FB3, 2020]. Any such action starts with identifying misinformation. Not everyone, however, agrees about which particular news items contain misinformation.

Journalists and professional fact checkers have the greatest expertise in assessing misinformation. Fact checkers have well-codified processes and criteria focusing on specific factual claims. On the other hand, typical news articles contain multiple factual claims, some of them more prominent than others. Thus, there is some inherent subjectivity in assessing how misleading an entire article is. Concerns have also been raised that professional fact checkers may skew liberal in their subjective judgments, despite their professional commitment to set aside personal biases.

Moreover, there is always some noise in any human judgment process, based on fatigue, distractions, and differences between raters. That noise can be reduced by averaging judgments from several people. It is expensive, however, to assemble large panels of journalists to assess news articles. Crowdsourcing, or using pools of laypeople, has been shown to be effective for many labeling tasks (e.g., Mitra and Gilbert [2015], Chung et al. [2019]). Social media platforms, however, have been reluctant to rely on lay raters to identify misinformation.

One concern is that lay raters will make judgments based on “gut reactions” to surface features, rather than determining the accuracy of factual claims. After all, the reason that misinformation circulates widely on social media platforms is that users upvote and share it, often without considering it deeply, or even reading it [Pennycook and Rand, 2019a], which is exacerbated by social media encouraging engagement more than accuracy [Jahanbakhsh et al., 2021]. There is reason, however, to be optimistic that this can be overcome. When asked to rate whether an article contains misinformation, people may assess it differently than they do when they choose whether to click or share an item. By simply priming
people to focus on accuracy, people’s intention to share other articles became more correlated with the accuracy of those articles [Pennycook et al., 2019].

Even when going beyond gut reactions, lay raters may lack the expertise to assess information in the way journalists do. Wineburg and McGrew [2019] found that journalists assessing an article tend to search for external corroborating or contradictory information, including what the authors call “lateral reading” about the author or source as well as the contents. College students and even history professors, by contrast, tended to focus more on signals within an article itself, to the detriment of their assessments. Here, too, however, there is reason for optimism. In another study, they found that with two 75-minute training sessions on searching for and interpreting external signals, students’ search practices and reasoning processes improved [McGrew et al., 2019].

A third concern is that lay raters may be ideologically motivated or biased. There are partisan divides in beliefs about some factual claims that have become politicized, such as the size of crowds at President Trump’s inauguration or whether wearing face masks significantly reduces the spread of COVID-19. Again, there is some reason for optimism that the problem of lay raters’ partisanship is not insurmountable. Recent work demonstrates that Democratic and Republican crowd workers largely agree when distinguishing mainstream news sites from hyper-partisan and fake news sites [Pennycook and Rand, 2019b]. It is unclear, however, what will happen when assessing individual articles, rather than news sources.

Two recent studies elicited misinformation judgments from lay raters about specific articles [Allen et al., 2021, Godel et al., 2021]. Perhaps surprisingly, they came to quite different conclusions. One found that a simple average of the ratings of a group of lay raters could identify misinformation pretty well, even though the raters saw just the headlines and ledes of articles and not the entire articles [Allen et al., 2021]. The other found that crowds performed worse than a journalist, even if fancy machine learning algorithms were used to aggregate the signals from the crowd members [Godel et al., 2021].

Here we report on a larger study, using articles from both of the other studies. More importantly, we study the effects of varying the elicitation process. In a control condition, raters assessed whether articles were false or misleading after opening the articles but without doing any additional research. In one treatment condition, each rater also searched for corroborating evidence. This was intended to elicit “informed” judgments rather than gut reactions. In the other treatment condition, raters consumed the results of others’ searches. This was intended to reduce ideological polarization—we deliberately included corroborating evidence links discovered by both liberal and conservative raters in an effort to broaden the search horizon for any given rater. Both of the treatment conditions thus enforced some form of lateral reading.

We benchmarked the performance of simulated subsets of our raters against simulated subsets of four journalists who also rated the articles. To enable an apples-to-apples comparisons, all simulated subsets of raters, whether lay raters or journalists, were scored by correlating their mean ratings with those of a single held-out journalist. We find that a large ideologically balanced panel of MTurk raters in the no research condition performed better than a benchmark of a single journalist. In the individual research condition, a large panel of lay raters performed better than a benchmark of a three-journalist panel. We also find that ideological polarization between liberal and conservative lay raters was reduced in both the individual and collective research conditions.

1 Experiment Design

We conducted a between-subjects study, where each lay rater only experienced one condition, but could rate as many articles as they wanted. Subjects were recruited and completed the labeling tasks through MTurk.

**Condition 1: No research.** The control condition asks participants to visit a URL (by clicking on a link in the interface) and then rate how “misleading” it was on a scale from 1=not misleading at all to 7=false or extremely misleading.

**Condition 2: Individual research.** The second condition adds an additional task before rating: search for supporting or challenging evidence, and paste a link to that evidence into the rating form. Otherwise, it was identical to the first condition.

**Condition 3: Collective research.** In the third condition the rater reviews links to corroborating or challenging information that raters from Condition 2 submitted, rather than searching for their own. We posted articles for Condition 3 after all Condition 2 ratings were received. For each article, we selected up to four links that were submitted by Condition 2 raters: the most popular link among liberals, the most popular among moderates, the most popular among conservatives, and the most popular overall. Often the same link was popular in more than one group, leading to fewer than four total URLs.
I.1 The Rating Apparatus

The second author conducted iterative usability studies on the survey software itself and the qualification tasks over the course of several months, with pilot participants thinking aloud while using the labeling software. Insights into question wording, presentation and order, and interface controls led to many modifications before data collection began. Figure 3 in the Appendix shows the labeling interface in the no research condition. Figure 4 in the Appendix shows the additional request made of raters in the second condition, to search for evidence and paste the search terms used and a link to the best evidence found. Figure 5 in the Appendix shows the interface in the third condition, where subjects were asked to click on links found by subjects in the second condition and select the one they thought provide the best evidence.

I.2 News Articles

Raters labeled two collections of English language news articles, taken from two other studies that were conducted independently in parallel with this one, by different teams. The first collection was selected from among 796 articles provided by Facebook that were flagged by their internal algorithms as potentially benefitting from fact-checking. Among those, 207 articles were manually selected based on the headline or lede including a factual claim, because the study’s focus was how little information needs to be presented to raters in order for them to make good judgments, and subjects were presented with only the headline and lede rather than viewing the entire article Allen et al. [2021]. Facebook also provided a topical classification based on an internal algorithm; of the 207 articles, 109 articles were classified by Facebook as “political”.

The second collection, containing 165 articles, comes from a study that focused on whether raters could judge items soon after they were published Godel et al. [2021]. It consisted of the most popular article each day from each of five categories: liberal mainstream news; conservative mainstream news; liberal low-quality news; conservative low-quality news; and low-quality news sites with no clear political orientation. Five articles per day were selected on 31 days between November 13, 2019 and February 6, 2020.

Our rating process occurred some months later. Four articles from the first collection and two articles from the second collection were removed because the URLs were no longer reachable when our journalists rated them, leaving a total of 368 articles between the two collections.

I.3 Crowds of Lay Raters

MTurk raters first completed a qualification task. Each was randomly assigned to one of three experiment conditions. They used the interface corresponding to their assigned experiment condition to label a sample item, after which they were given two attempts to correctly answer a set of multiple choice questions quizzing them on their understanding of the instructions. They then completed an online consent form, a four-question multiple choice knowledge quiz, and a questionnaire about demographics and ideology. Subjects who did not pass the quiz about the instructions, or answer at least two questions correctly on the political knowledge quiz, were excluded. Table 1 shows the recruitment funnel. Because the individual research condition had slightly more complex instructions, there were more opportunities to fail the instructions quiz and thus there was a higher rate of failure. In order to get equal numbers of subjects who passed the qualification test, more subjects had to be randomized to the individual research condition. In the discussion section, we return to the question about whether the results may be due to selection of a more qualified or conscientious pool of raters. Of subjects who completed the qualification, a higher percentage of subjects in the second condition went on to complete rating tasks.

At the completion of the qualification task, workers were assigned to one of nine groups based on their randomly assigned condition and their ideology. We asked about both party affiliation and ideology, each on a five-point scale; raters who both leaned liberal and leaned toward the Democratic Party were classified as liberal; those who both leaned conservative and leaned toward the Republican Party were classified as conservative; others were classified as moderate. In each of the three conditions, eighteen people of each ideology rated each news article.

Table 2 shows that more raters were liberal than moderate or conservative. Thus, in order to get eighteen ratings from each rater group, the moderate and conservative raters rated more items on average than the liberal raters, as shown in Table 3.

In the published report, they analyzed results for only 99 of the 165 articles, excluding those from liberal and conservative mainstream sites. They also added an additional 36 articles in a second wave conducted that were not assessed in our study.
INFORMED CROWDS CAN EFFECTIVELY IDENTIFY MISINFORMATION

Table 1: Recruitment funnel

|                          | C1: No Research | C2: Individual Research | C3: Collective Research |
|--------------------------|-----------------|-------------------------|-------------------------|
| randomized               | 1530            | 2185                    | 1583                    |
| passed qualification     | 623             | 622                     | 622                     |
| rated at least 1 item    | 404             | 500                     | 397                     |
| median ratings per person| 14              | 14                      | 14                      |
| mean ratings per person  | 49.6            | 40.1                    | 50.3                    |
| max ratings by one person| 352             | 368                     | 366                     |

Table 2: Raters who submitted at least one rating, by condition and ideology

|                  | C1: No Research | C2: Individual Research | C3: Collective Research |
|------------------|-----------------|-------------------------|-------------------------|
| Liberal          | 184             | 247                     | 177                     |
| Moderate         | 113             | 107                     | 123                     |
| Conservative     | 107             | 146                     | 97                      |

Raters had the option to say they did not have enough information to make a judgment. This option was rarely used (<3% in all conditions, see Table 5 in the appendix). Such ratings were excluded when computing averages. Figures 6 and 7 in the appendix show the frequency of judgments, by treatment condition and by ideology.

1.4 Journalist Raters

Four journalists rated all items from both collections. They used the labeling interface for Condition 2, which required them to do individual research. All four had just completed a prestigious, selective fellowship for mid-career journalists. Three were U.S.-based and the fourth had covered U.S. politics for many years. One had been through the American Press Institute’s fact-checking bootcamp. One journalist did not rate one item. Another journalist did not rate two items and reported that they did not have enough information to make a judgment on 54 others. Such missing ratings were excluded when computing averages and correlations.

2 Measures and Results

We report three evaluation metrics. First is inter-rater agreement, a measure of internal consistency. Second is ideological polarization, measured by correlation between liberal and conservative ratings. Third is correlation with journalist ratings, which we compare to benchmarks of journalist to journalist correlation.

2.1 Inter-Rater Agreement

Figure 1 provides a heatmap of the frequency of the 1-7 ratings for each item in the three conditions. Each item is a row and rows are sorted based on the mean ratings for the item across all three conditions. Inter-rater agreement summary statistics are computed from these data. The color coding makes it obvious that in the two research conditions, there

Table 3: Median ratings per Turker

|                | C1: No Research | C2: Individual Research | C3: Collective Research |
|----------------|-----------------|-------------------------|-------------------------|
| Liberal        | 10              | 11                      | 13                      |
| Moderate       | 20              | 20                      | 14                      |
| Conservative   | 15              | 17                      | 16                      |
INFORMED CROWDS CAN EFFECTIVELY IDENTIFY MISINFORMATION

Figure 1: Frequency of the 1–7 ratings for each item in the three conditions

were many more items with a consensus of 1 (not misleading at all) or 7 (false or extremely misleading), and fewer intermediate ratings.

As a summary measure of agreement, we report the intraclass correlation of ratings (ICC) \cite{Shrout and Fleiss 1979}. Row 1 of Table 4 shows the ICC estimates. We find that the ICC is highest in the collective research condition and that the differences are statistically significant.

2.2 Agreement between Partisan Groups

We also assess whether there is systematic disagreement between raters with different political ideologies. For each item, we compute the mean rating among the eighteen liberal raters and the mean rating among the conservative raters. We then compute the Pearson correlation coefficient, across items, of these mean ratings. The liberal-conservative correlation was .82 in the no-research condition, .88 in the individual research condition and .90 in the collective research condition.

The effects, however, were not uniform for all kinds of items. On just the political items from the first collection, the liberal-conservative correlations in the three conditions were .69, .81, and .79. On just the non-political items from that collection, the correlations were .90, .93, and .98. On the items from the second collection, the liberal-conservative agreement hardly varied between conditions: .83, .83, and .85. See the Appendix for details about results on subsets of the items.

2.3 Agreement with Journalists

Our external validity metric is to compare the performance of panels of lay raters to the performance of benchmark panels of journalists. Journalists make an appropriate benchmark because of their expertise and professional ethos.

To enable performance comparison, both lay rater panels and journalist panels are scored across the whole set of items based on their correlation with the ratings of a single held-out journalist. In other words, we measure the ability of the panel to predict the ratings of a held-out journalist. The logic is that each single journalist is presumed to be reasonably correlated with the underlying, hidden ground truth. Thus, if one panel correlates with a single journalist better than another panel does, we infer that the first panel correlates better with the hidden ground truth.
Benchmarking against the performance of simulated panels of journalists is especially interesting because platforms may face the practical question of whether and when to rely on crowd judgments to extend the reach of journalists’ or fact checkers’ judgments which are available for only a limited set of items. If a platform would be willing to rely on the judgment of a single journalist, then arguably it should be willing to rely on a crowd that performs better than a single journalist at predicting what another journalist would say. If a platform would be willing to rely on the majority rating of three journalists, then arguably it should be willing to rely on a crowd that performs better than a panel of three journalists at predicting what another journalist would say.

Note that we always score simulated panels against a single held-out journalist. If, by contrast, we scored panels against the mean of several journalists rather than against a single journalist, the correlation scores for all simulated panels would be higher, but the relative ordering of scores should not change.

By scoring all panels against just a single journalist, we leave three of the four journalists available for inclusion in simulated benchmark panels. Thus, we are able to compare the performance of large lay panels to the performance of groups of two and three journalists. Considering all four possible panels of three journalists, the average correlation with the remaining journalist was 0.73. For panels of two journalists, the average correlation with a held-out journalist was 0.70. For the twelve possible pairs of journalists, the average correlation with a single held-out journalist was 0.65.

Following the procedure in Resnick et al. [2021], we construct simulated panels of lay raters of various sizes, from one to fifty-four, to produce a “power curve”, as shown in Figure 2. Each point on the power curve is the expected correlation of a randomly selected group of lay raters of that size with a randomly selected journalist, the power of a panel of that size to predict what a journalist will say. We use a bootstrap procedure to estimate confidence intervals over 500 resampled sets of items.

The intersection points in the graphs in Figure 2 show the number of lay raters required to get the same predictive power as one or three journalists. For example, in the no-research condition, 7.69 lay raters were sufficient to achieve the same power as one journalist.

Comparing across the three conditions, we can see that lay raters in the two research conditions correlate with a journalist better than do raters in the no research condition and that the individual research condition has greater power than the collective research condition for large groups of lay raters. In the individual research condition, C2, 15.22 lay raters were equivalent to three journalists; even 54 raters were not sufficient to achieve the same power as three journalists in the other two conditions, and in the no research condition, C1, 54 raters were also insufficient to achieve the same power as two journalists.

To assess reliability we considered the results separately for each bootstrap sample of items (see Table 4 in the Appendix). For all group sizes, the power of a group of lay raters in the individual research condition was higher than one in the no research condition.

Table 4: Summary of results by experimental condition.

| Condition     | ICC  | C2   | C3   |
|---------------|------|------|------|
| Liberal-Conservative Correlation | 0.82 | 0.88 | 0.90 |
| Correlation of one lay rater with one journalist | 0.46 | 0.52 | 0.52 |
| Lay raters equivalent to one journalist | 7.69 | 3.40 | 3.80 |
| Lay raters equivalent to two journalists | >54  | 7.08 | 13.55|
| Lay raters equivalent to three journalists | >54  | 15.22| >54  |
| Liberal raters equivalent to one journalist | 3.54 | 2.13 | 2.20 |
| Liberal raters equivalent to two journalists | 10.42| 3.91 | 5.11 |
| Liberal raters equivalent to three journalists | >18  | 6.82 | 13.81|
| Conservative raters equivalent to one journalist | >18  | 5.48 | 9.68 |
| Conservative raters equivalent to two journalists | >18  | >18  | >18  |
| Conservative raters equivalent to three journalists | >18  | >18  | >18  |

2 Non-integer equivalence values can be interpreted as a randomization procedure. For example, 7.69 lay raters corresponds to using eight raters for 69% of the items, and seven raters on the other items.
research condition in all of the 500 bootstrap samples. With just a single rater, the collective research condition had higher power than the individual research condition on 71.5% of the item samples. With groups of five raters, however, the individual research condition performed better than the collective research condition on 98.4% of item samples.

In condition C2, with individual research, eighteen lay raters outperformed two journalists on all of the bootstrap item samples and outperformed three journalists on 65.7% of samples (see Table 8 in the Appendix). 54 lay raters outperformed three journalists on 97.4% of bootstrap item samples.

3 Discussion

3.1 Informed Raters Make Better Judgments

As noted in the introduction, researchers found that with two 75-minute training sessions, students adopted practices of external search for supporting and challenging information, and lateral reading about the author and source Wineburg and McGrew [2019]. In our study, none of the raters received explicit training. Both of the treatment conditions, however, required raters to seek out or consider external evidence. This led to improved misinformation judgments, as measured in a variety of ways. Judgments in conditions 2 and 3 were more internally consistent between raters, showed less partisan divide, and were better correlated with expert journalist judgments.

3.2 How to Best Elicit Misinformation Judgments

The results are mixed about what is the best way to ask raters to consider external evidence. Condition 2, where each person searched individually, had lower consistency among raters than Condition 3 and more partisanship. However, when averaging ratings from several raters, the correlation with a journalist was higher in Condition 2, with the difference becoming more and more reliable when averaging across larger rater sets. Each person doing their own search seems to yield a little more information from each person, but also more noise in individual assessments. Wisdom of crowds models posit that when averaging several judgments it is better for those judgments to be independent Surowiecki [2005]. Our results are consistent with that; in Condition 3, examining a common set of links to potential corroborating or challenging evidence may have yielded correlated errors in judgment. We suspect that the best procedure for eliciting judgments from raters will be some hybrid that encourages both individual search and considering links that have been discovered by others, especially links discovered by people with different ideologies.

If one imagines that the primary use of crowd-based misinformation judgments will be to judge hundreds or thousands of articles every day, basically to “scale up” real-time fact-checking, then cost-effectiveness is an important consideration. An alternate vision, which we favor, conceives of using citizen juries as a governance mechanism, rather than as a way to speed up decision making Zittrain [2019], Fan and Zhang [2020]. In that vision, crowd-based misinformation judgments could be used as as part of appeals processes, as ground truth for transparency reports about platform performance, and
as training data for human and automated processes. For that, they would need to operate on a medium rather than large scale.

In either case, we think that it is most important at this stage in the development of crowd-sourced misinformation judgments to focus on processes that maximize quality rather than minimizing costs; cost optimization can come a little later. Even so, it is interesting to consider the cost effectiveness of different ways of eliciting judgments.

We offered raters in the no research condition $.50 for each article rated and in the two research conditions $1 for each article rated, in order to compensate them for the extra time required to do research. If, in practice, one’s goal was to get a rating of quality similar to what one would get from a single journalist, this could have been accomplished for $5.50 using eleven raters in the no research condition or $4 using four raters in either of the research conditions.

3.3 Limitations

One limitation of our study is that Turkers passed the qualification test at a lower rate in Condition 2. Thus, it is possible that the better performance in that condition was due, in part, to a pool of raters who were more diligent or skilled, rather than the requirement that they search for a corroborating source. It would be interesting in a future study to tease apart the selection effect from the task effect; if the selection effect is sufficient to yield the rater performance found in the second condition, without requiring raters to actually perform independent research, the costs of rater labeling could be further reduced.

Another limitation is that articles were rated after they were first posted. It is possible that searching for corroborating evidence would not be as impactful soon after the articles were posted, and thus not as effective at driving raters to provide better misinformation judgments. The study that generated our second collection of articles was explicitly designed to compare judgments made by lay raters and journalists within a few days of an article’s publication. It would be interesting to analyze how well those ratings correlate with our journalist and lay ratings that were collected several months later.

Finally, more training for journalists, more time spent evaluating each article, or incentives for agreeing with each other could lead to higher inter-rater agreement among journalists. That would set a higher benchmark for lay panels to compete against, as noted in [Godel et al. 2021].

3.4 Reconciling Results with Previous Studies

It is worth revisiting the two previous studies to try to assess possible reasons for discrepant results. Because we reused most of the same items, we can rule out some possible explanations, while others will require further research to tease apart.

First, we note that the three studies constructed the journalist benchmark slightly differently. All three studies compared a simulated panel of lay raters to a benchmark simulated panel of one or more journalists. We ensured apples-to-apples comparisons by scoring both lay panels and journalist panels ability to predict a single held-out journalist’s ratings. However, for scoring the single journalist, they had to exclude that journalist, and thus they scored the journalist against the majority vote of a smaller panel. Most likely, the majority vote of a smaller panel will have higher variance, which would make their single journalist benchmark score be lower than it would be in a fair comparison. However, because of the tie-breaking methods they used (a 2-2 vote of four journalists was treated as “not misinformation”), it is possible that comparing to a smaller panel may have produced a score that was higher than it would be in a fair comparison.

Another possible explanation for discrepant results is uncertainty due to sampling error. Godel et al. [2021] reported on one binary outcome (whether the panel’s prediction matched the majority vote of the journalists) for each of 135 articles. The accuracy reported for the majority vote of a random crowd of 25 people was 62%, and for the single journalist benchmark it was 69%. But this difference in performance could easily occur by chance even if there was no difference between the two. Suppose, for example, that panels of 25 people and the benchmark journalist both had 65% prediction
Informed Crowds Can Effectively Identify Misinformation

accuracy. The 95% confidence interval for the sampling distribution of 135 draws from a Bernoulli distribution with
p=0.65 is ±0.08.

In our study, we have a larger pool of items. We quantify the uncertainty of our results through bootstrap sampling
of 500 simulated article sets. In our individual research condition, panels of seven or more lay raters outperformed a
benchmark of a single journalist on all 500 simulated article sets and the full panel of 54 lay raters outperformed a
benchmark of three journalists on 98% of the simulated article sets.

An even bigger source of uncertainty is the individual journalists. We employed just four journalists and the other
studies employed three and six. Even one or two who were outliers from their peers could have reduced the average
ability of benchmark journalist panels to predict what other benchmark journalists would say. Since we included the
same items from [Allen et al., 2021] in our study, both teams were able to perform robustness checks using the other
team’s journalist ratings. Tables 9 and 10 in the Appendix show qualitatively similar results in our study whether we
use ratings from our four journalists or from their three journalists for the analysis. Because the journalist ratings from
the other study are not available, we have not been able to perform a similar robustness check on the items from the
second article collection.

A third possible reason for discrepant results could be differences between the types of items in the two sets. The first
article collection was selected to include only articles that included a factual claim in the headline. These might be
easier for lay raters to judge. To assess this, we ran our analyses separately on the two article collections. Results
were qualitatively similar for both. On the 98 articles from the second collection that were from fringe rather than
mainstream sources, the subset which were analyzed in [Godel et al., 2021], lay raters correlated with a journalist less
well but journalists also correlated with each other less well; nineteen or more lay raters outperformed a benchmark
of three journalists (see Tables 11 and 12 in the Appendix). We did, however, find that lay rater performance may
have been worse on the subset of 109 political items from the first collection, with panels of 15.26 raters matching
the performance of two journalists but even the full panel of 54 lay raters failing to match the performance of three
journalists, (see Table 13 and Figure 9 in the Appendix). Because of the smaller sample size, however, confidence
intervals are wider, so there is uncertainty about whether performance on these kinds of items is truly different.

[Godel et al., 2021] suggests that the timing of when rating happens may be critical to the capabilities of lay crowds.
In particular, they collected ratings within three days after articles first appeared. For some news items, follow-up
articles or fact-checks might appear later that help to disseminate correct information. Prior to that, lay raters may be
poor judges of information quality. Our study and Allen et al. [2021] had both lay raters and journalists assess articles
weeks to months after they first appeared. If lay crowds are especially poor at assessing articles soon after they are
posted, relative to a benchmark of journalists, that would limit their use for real-time decision-making. As mentioned
previously, if lay panels are conceived of as citizen juries, as part of platform governance and transparency procedures,
the ability to make real-time judgments may not be as important.

Finally, we note significant differences in the rating procedures between the studies. Our finding that more informed
raters make better judgments may explain some of the differences in results. The study from which we drew the first
collection of articles asked raters to examine only the headline and lede, and in one condition the source, without
clicking through to see the whole article [Allen et al., 2021]. As noted, they found that the average of fifty lay raters had
slightly lower performance than one fact-checker. Our first condition provided a little more information to raters, by
asking them to examine the full article, and panels of eight or more lay raters had a higher performance than a single
journalist. Our second and third conditions, involving individual research or reviewing the results of collective research,
provided our raters with even more information, and led to still better judgments.

It is also worth noting that refinement of the selection process and the user interface and instructions may make a
big difference for the performance of lay crowds. Studies of crowdsourced labeling in other domains have found that
quality control measures and small amounts of training are helpful [Mitra et al., 2015]. For this study, we developed a
custom web interface embedded as an iframe within MTurk pages, and went through multiple rounds of UX testing over
several months. We also excluded raters who did not pass a simple test of whether they understood the interface and
instructions and those who did not correctly answer two out of four knowledge questions. Finally, many MTurk workers
are very conscientious, especially if they fear work rejections that will harm their ability to earn money in the future.
In this study, in addition to the misinformation judgments, workers were asked to provide a subjective opinion about what
enforcement action they thought platforms should take and to make a prediction about other raters’ subjective opinions.
They were told that their judgments and opinions would not be evaluated, but that if their predictions were too far off
they would be disqualified, and this may have encouraged workers to be conscientious. All of these factors may have
helped the lay crowds in our study to perform better than they might have otherwise.
3.5 A Model of Crowdsourced Judgments

Consider a stylized hierarchical model of individual raters’ judgments when those raters are drawn from some defined pool (e.g., journalists, or liberal workers on Mechanical Turk):

\[ \text{Judgment} \sim \text{Truth} + \text{GroupOffset} + \text{RaterNoise} \]

*Truth* for an article is an unobserved hypothetical “correct label” for an item. The *GroupOffset* for an article is the difference between the Truth and the (also unobserved) mean rating in the rater population. *Noise* is the deviations of individual raters from that rater population’s mean for an item, due to inter-individual differences and factors like fatigue and distractions.

In this model, there is a Journalist Consensus for each item, consisting of the Truth plus possibly a GroupOffset. And then each journalist sees a draw from a distribution centered on the Journalist Consensus for that item. Similarly, in each condition, for each item there is a Liberal Lay Rater Consensus, a Conservative Lay Rater Consensus, and an overall Lay Rater Consensus.

The GroupOffset can be further decomposed into two components. One is due to *imperfect expertise* of the rater pool and the other due to *ideological bias*.

While none of the underlying components, Truth, GroupOffset, and RaterNoise, can be directly observed, this model provides a framework for interpreting many of the results of our study. Most importantly, the relatively high absolute correlations among judgments from random pairs of raters in all conditions suggest that the Truth component of judgments was fairly large; perhaps the notion of truth is not completely broken beyond repair.

The noise in individual ratings was much higher for lay raters than for journalists. This can be inferred from the correlations between random pairs of raters from each group. The average correlation between a pair of random lay raters in the individual research condition was 0.47. For a random pair of journalists, it was 0.65.

The intuition behind the so-called wisdom of crowds is that, following the central limit theorem, the mean of a large sample of independent draws from a distribution has lower variance than the mean of a single draw [Surowiecki, 2005]. In our model, taking the mean of many raters should reduce the RaterNoise. We find just that: the correlation between the means of random collections of eighteen raters in the individual research condition was 0.94, up from 0.47 for pairs of individual raters.

In our study, the reduction in noise was sufficient to overcome the presumed smaller GroupOffset for journalists because of their expertise and professionalism. To see this, note that the correlation of a journalist with the mean of a group of eighteen lay raters in the individual research condition (C2) was .74, much better than the 0.65 correlation between a journalist and another journalist.

Liberal and conservative lay raters had systematically different GroupOffsets. This follows from the fact that groups of liberal and conservative lay raters correlated with each other less than groups of random lay raters. Even in the independent research condition (C2), where this effect was reduced, the correlation between the means of eighteen liberal and eighteen conservative lay raters was 0.88, short of the 0.94 correlation between the means of random collections of eighteen lay raters. This could reflect ideological bias in judgments by one or both groups, or differences in expertise between the groups. We cannot distinguish between ideological bias and expertise differences, nor can we determine which group’s Consensus tended to be closer to the hypothetical Truth.

Consequently, we also cannot determine definitively whether the journalists had any bias in their judgments. Liberal lay raters correlated better with journalists than conservative lay raters did, meaning that the Journalist Consensus tended to be closer to the Liberal Lay Rater Consensus than it was to the Conservative Lay Rater Consensus. To the extent that liberal lay raters had ideological bias contributing to their GroupOffsets, the GroupOffsets for journalists must reflect similar ideological bias. On the other hand, if the GroupOffsets for liberals reflected only limited expertise, then a better correlation with journalists might not imply any ideological bias on the part of the journalists.

4 Conclusion

Overall, the results suggest that juries composed of lay raters, perhaps the users of social media platforms themselves, could be a valuable resource in assessing misinformation. Together, their assessments can be comparable to or better than asking three expert journalists to make assessments. Importantly, these pools of raters should be readily available to most large social media platforms. Moreover, they could be selected to deliberately be representative of different ideological viewpoints, thus reducing concerns about potential bias of raters who are selected solely for their expertise.

Much work, however, to refine the processes by which rater pools are selected and the research tasks they are asked to perform. Would even more research, or more research done in a different way, further reduce ideological...
disagreement between liberals and conservatives and further increase alignment with journalist ratings? Moreover, the present work treats all disagreements in judgments as equally important. However, in reality, we know that certain disagreements (e.g., election fraud, the dangers of COVID-19, etc.) have outsized capability to harm society. More work should be done to systematically understand these factors and how to address them.

5 Acknowledgments

We thank the four journalists who evaluated the articles. We thank the two other study teams whose article collections we used in this paper. In addition to sharing the article URLs, they provided useful feedback at a workshop where study designs were shared and at a second where preliminary results were shared. In particular, we thank Kevin Aslett, Adam Berinsky, William Godel, Amber Heffernan, Quentin Hsu, Jenna Koopman, Nathan Persily, David Rand, Zeve Sanderson, Luis Sarmenta, Henry Silverman, and Josh Tucker. We thank Kelly Garrett and Eshwar Chandrasekharan for early conversations that influenced the design of the study. Both workshops were sponsored by Facebook. Facebook also provided funding for this study. One of the authors of this study, (Paul Resnick), served as a paid consultant to Facebook from May 2018-May 2020.

References

A Mitchell, M Jurkowitz, J Baxter Oliphant, and E Shearer. Americans who mainly get their news on social media are less engaged, less knowledgeable. *Pew Research Center: Journalism & Media*, 2020.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

Facebook’s third-party fact-checking program. [https://web.archive.org/web/20200603111446/https://www.facebook.com/journalismproject/programs/third-party-fact-checking](https://web.archive.org/web/20200603111446/https://www.facebook.com/journalismproject/programs/third-party-fact-checking), 2020.

Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pages 258–267, 2015.

John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

Gordon Pennycook and David G Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019a.

Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021. doi:10.1145/3449092. URL [https://doi.org/10.1145/3449092](https://doi.org/10.1145/3449092).

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David Rand. Understanding and reducing the spread of misinformation online. *Unpublished manuscript: https://psyarxiv.com/3n9u8*, 2019.

Sam Wineburg and Sarah McGrew. Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, 121(11):1–40, 2019.

Sarah McGrew, Mark Smith, Joel Breakestone, Teresa Ortega, and Sam Wineburg. Improving university students’ web savvy: An intervention study. *British Journal of Educational Psychology*, 89(3):485–500, 2019.

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

Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eaba4393, 2021. doi:10.1126/sciadv.ab4393. URL [https://www.science.org/doi/abs/10.1126/sciadv.ab4393](https://www.science.org/doi/abs/10.1126/sciadv.ab4393).

William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, 1(1), 2021.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.
6 Appendix

6.1 ICC computation

The ICC is computed with a one-way random effects model, using the mean of raters and consistency as internal metrics \cite{Koo2016}; we compute the ICC using the rpy2 bridge library \cite{Rpy2}, which accesses an underlying CRAN implementation of the ICC \cite{ICCPdf}. This formulation of the ICC accounts for unstable rater pools: i.e., there is no guarantee that the same group of raters rates each URL.

We compute the overall ICC and confidence intervals of each condition using a bootstrap procedure: we simulate 10,000 draws from our data via sampling with replacement.

The 95% confidence intervals for the three conditions are: [0.407, 0.452], [0.449, 0.481], and [0.502, 0.529].

6.2 Reliability of Comparisons Between Conditions

For each of our 500 bootstrap samples of items, we computed the power curve giving the average, over many samples of k turkers, of the correlation of the mean of those turker ratings with a journalist’s ratings. This was done in each condition. As a measure of the reliability of comparisons between conditions, we report the fraction of those samples where the correlation was higher between one condition than another, as shown in Table 6.

6.3 Reliability of Survey Equivalence Values

For each of our 500 bootstrap samples of items, we computed the power curve giving the average, over many samples of k turkers, of the correlation of the mean of those turker ratings with a journalist’s ratings. From this a survey-equivalence value was computed: how many turkers are equivalent to one, two, or three journalists. In Table 7 we report a range that includes 95% of the values computed for the bootstrap samples. In Table 8 we report the fraction of bootstrap samples where k lay raters outperformed m journalists in predicting the ratings of a held-out journalist.

6.4 Results for Item and Rater Subsets

As described in the main text, there were 207 news items in the first sample, of which 109 were marked as political. There were 165 items in the second sample.

We have ratings from four journalists for both samples. In addition, the team that assembled the first sample of items collected ratings from three other journalists, as described in \cite{Allen2021}. The journalists did not answer exactly the same question we posed, but were also on a 1-7 scale. We were able to treat them as comparable to our journalist ratings after reverse coding them.

\footnote{\url{https://rpy2.bitbucket.io}}
\footnote{\url{https://cran.r-project.org/web/packages/ICC/ICC.pdf}}
In Tables 9 and 11 we report key result metrics for the two article collections separately. Table 10 reports key result metrics for the three journalists from Allen et al. [2021]. Table 13 reports key result metrics for the only the 109 political items.

Since the results are somewhat different for the political items, Figure 9 plots the power curves and survey equivalence values for just those items.
Read
First, open the news item. **Skim** it for **about one minute**.

**News Item to Assess** (click to open in new tab)

Evidence
Second, take **up to five minutes** to search, using a search engine, for evidence that will help you judge the news item. You should look for both **supporting and challenging** evidence.

What search terms did you use?

Did you find any page with evidence you found convincing, one way or the other?

- Yes
- No

If Yes, please paste the link you found most relevant and convincing.

**Figure 4**: Labeling interface for Condition 2: Independent Research.

Read
First, open the news item. **Skim** it for **about one minute**.

**News Item to Assess** (click to open in new tab)

Evidence
Second, which of the following links provides the most convincing evidence either supporting or contradicting claims in the original news item?

- Evidence link 1
- Evidence link 2
- Evidence link 3
- None of the above provides useful evidence

**Figure 5**: Labeling interface for Condition 3: Collective Research.
Table 5: Percent of Raters Who Reported Not Enough Information to Make a Judgment, per Item

|                  | C1: No Research | C2: Individual Research | C3: Collective Research |
|------------------|-----------------|--------------------------|-------------------------|
| Mean             | 1.5%            | 3.0%                     | 1.4%                    |
| Median           | 0               | 1.9%                     | 0                       |
| Max              | 7.4%            | 22.2%                    | 35.7%                   |
Figure 6: Frequency of Judgments By Condition
Figure 7: Frequency of Judgments By Ideology
Table 6: Reliability of Comparisons Between Conditions

| number of lay raters | C2 > C1 | C3 > C1 | C3 > C2 |
|---------------------|---------|---------|---------|
| 1                   | 1.00    | 1.00    | 0.71    |
| 2                   | 1.00    | 1.00    | 0.28    |
| 3                   | 1.00    | 1.00    | 0.24    |
| 4                   | 1.00    | 1.00    | 0.05    |
| 5                   | 1.00    | 1.00    | 0.02    |
| 6                   | 1.00    | 1.00    | 0.00    |
| 7                   | 1.00    | 1.00    | 0.00    |
| 8                   | 1.00    | 0.99    | 0.01    |
| 9                   | 1.00    | 0.98    | 0.00    |
| 10                  | 1.00    | 0.97    | 0.00    |
| 12                  | 1.00    | 0.98    | 0.00    |
| 14                  | 1.00    | 0.98    | 0.00    |
| 16                  | 1.00    | 0.96    | 0.00    |
| 18                  | 1.00    | 0.96    | 0.00    |
| 20                  | 1.00    | 0.96    | 0.00    |
| 22                  | 1.00    | 0.95    | 0.00    |
| 24                  | 1.00    | 0.96    | 0.00    |
| 26                  | 1.00    | 0.95    | 0.00    |
| 28                  | 1.00    | 0.95    | 0.00    |
| 30                  | 1.00    | 0.95    | 0.00    |
| 32                  | 1.00    | 0.94    | 0.00    |
| 34                  | 1.00    | 0.95    | 0.00    |
| 36                  | 1.00    | 0.94    | 0.00    |
| 38                  | 1.00    | 0.94    | 0.00    |
| 40                  | 1.00    | 0.94    | 0.00    |
| 42                  | 1.00    | 0.94    | 0.00    |
| 44                  | 1.00    | 0.94    | 0.00    |
| 46                  | 1.00    | 0.94    | 0.00    |
| 48                  | 1.00    | 0.94    | 0.00    |
| 50                  | 1.00    | 0.94    | 0.00    |
| 52                  | 1.00    | 0.94    | 0.00    |
| 54                  | 1.00    | 0.94    | 0.00    |
Table 7: Survey equivalence confidence intervals. How many lay raters provide the same power to predict a journal as a panel of m journalists? Confidence intervals based on covering 95% of the survey equivalence values from 500 bootstrap samples of articles.

| m journalists | C1: No Research | C2: Individual Research | C3: Collective Research |
|---------------|-----------------|--------------------------|------------------------|
| 1             | (3.92, 24.58)   | (2.53, 4.62)             | (2.62, 5.94)           |
| 2             | (12.26, >54)    | (4.75, 12.14)            | (6.01, >54)            |
| 3             | (>54, >54)      | (8.30, 53.42)            | (17.38, >54)           |
Table 8: Fraction of bootstrap samples of articles where k lay raters are better than m journalists (i.e., higher correlation with held-out journalist)

| k lay raters | C1: No Research | C2: Individual Research | C3: Collective Research |
|--------------|-----------------|-------------------------|-------------------------|
| C2 m journalists | C1 m journalists | C2 m journalists | C3 m journalists |
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3 | 0.00 | 0.00 | 0.00 | 0.24 | 0.00 | 0.00 | 0.16 | 0.00 | 0.00 |
| 4 | 0.03 | 0.00 | 0.00 | 0.88 | 0.00 | 0.00 | 0.63 | 0.00 | 0.00 |
| 5 | 0.09 | 0.00 | 0.00 | 1.00 | 0.06 | 0.00 | 0.89 | 0.01 | 0.00 |
| 6 | 0.19 | 0.00 | 0.00 | 1.00 | 0.29 | 0.00 | 0.98 | 0.03 | 0.00 |
| 7 | 0.38 | 0.00 | 0.00 | 1.00 | 0.50 | 0.00 | 0.99 | 0.08 | 0.00 |
| 8 | 0.58 | 0.00 | 0.00 | 1.00 | 0.68 | 0.01 | 1.00 | 0.18 | 0.00 |
| 9 | 0.67 | 0.00 | 0.00 | 1.00 | 0.83 | 0.05 | 1.00 | 0.21 | 0.00 |
| 10 | 0.76 | 0.01 | 0.00 | 1.00 | 0.91 | 0.13 | 1.00 | 0.28 | 0.00 |
| 12 | 0.82 | 0.02 | 0.00 | 1.00 | 0.97 | 0.26 | 1.00 | 0.41 | 0.01 |
| 14 | 0.89 | 0.04 | 0.00 | 1.00 | 0.99 | 0.44 | 1.00 | 0.57 | 0.01 |
| 16 | 0.93 | 0.06 | 0.00 | 1.00 | 1.00 | 0.57 | 1.00 | 0.61 | 0.02 |
| 18 | 0.95 | 0.11 | 0.00 | 1.00 | 1.00 | 0.66 | 1.00 | 0.68 | 0.03 |
| 20 | 0.96 | 0.11 | 0.00 | 1.00 | 1.00 | 0.72 | 1.00 | 0.73 | 0.04 |
| 22 | 0.97 | 0.14 | 0.00 | 1.00 | 1.00 | 0.77 | 1.00 | 0.76 | 0.06 |
| 24 | 0.97 | 0.15 | 0.00 | 1.00 | 1.00 | 0.80 | 1.00 | 0.80 | 0.08 |
| 26 | 0.98 | 0.19 | 0.00 | 1.00 | 1.00 | 0.83 | 1.00 | 0.83 | 0.09 |
| 28 | 0.98 | 0.22 | 0.00 | 1.00 | 1.00 | 0.85 | 1.00 | 0.83 | 0.10 |
| 30 | 0.98 | 0.25 | 0.00 | 1.00 | 1.00 | 0.89 | 1.00 | 0.86 | 0.11 |
| 32 | 0.98 | 0.27 | 0.01 | 1.00 | 1.00 | 0.91 | 1.00 | 0.87 | 0.12 |
| 34 | 0.98 | 0.27 | 0.01 | 1.00 | 1.00 | 0.92 | 1.00 | 0.89 | 0.12 |
| 36 | 0.99 | 0.30 | 0.01 | 1.00 | 1.00 | 0.94 | 1.00 | 0.90 | 0.13 |
| 38 | 0.99 | 0.32 | 0.01 | 1.00 | 1.00 | 0.95 | 1.00 | 0.90 | 0.14 |
| 40 | 0.99 | 0.32 | 0.01 | 1.00 | 1.00 | 0.96 | 1.00 | 0.90 | 0.14 |
| 42 | 0.99 | 0.33 | 0.01 | 1.00 | 1.00 | 0.96 | 1.00 | 0.91 | 0.15 |
| 44 | 0.99 | 0.34 | 0.01 | 1.00 | 1.00 | 0.96 | 1.00 | 0.91 | 0.15 |
| 46 | 0.99 | 0.35 | 0.01 | 1.00 | 1.00 | 0.96 | 1.00 | 0.91 | 0.16 |
| 48 | 0.99 | 0.36 | 0.01 | 1.00 | 1.00 | 0.96 | 1.00 | 0.92 | 0.17 |
| 50 | 0.99 | 0.37 | 0.01 | 1.00 | 1.00 | 0.97 | 1.00 | 0.92 | 0.17 |
| 52 | 0.99 | 0.37 | 0.01 | 1.00 | 1.00 | 0.97 | 1.00 | 0.93 | 0.18 |
| 54 | 0.99 | 0.39 | 0.01 | 1.00 | 1.00 | 0.97 | 1.00 | 0.93 | 0.19 |
|                          | C1: No Research | C2: Individual Research | C3: Collective Research |
|--------------------------|-----------------|-------------------------|-------------------------|
| Liberal-Conservative Correlation | .80             | .88                     | .91                     |
| Correlation of one lay rater with one journalist | 0.47            | 0.53                    | 0.54                    |
| Lay raters equivalent to one journalist | 7.85            | 3.80                    | 4.29                    |
| Lay raters equivalent to two journalists | >54             | 9.06                    | 25.92                   |
| Lay raters equivalent to three journalists | >54             | 25.11                   | >54                     |
| Liberal raters equivalent to one journalist | 3.86            | 2.48                    | 2.33                    |
| Liberal raters equivalent to two journalists | 13.29           | 5.00                    | 6.26                    |
| Liberal raters equivalent to three journalists | >18             | 10.01                   | >18                     |
| Conservative raters equivalent to one journalist | >18             | 6.78                    | >18                     |
| Conservative raters equivalent to two journalists | >18             | >18                     | >18                     |
| Conservative raters equivalent to three journalists | >18             | >18                     | >18                     |

Table 9: Summary of results for article collection one only, using our four journalist ratings.
|                              | C1: No Research | C2: Individual Research | C3: Collective Research |
|------------------------------|-----------------|-------------------------|------------------------|
| Liberal-Conservative Correlation | .80             | .88                     | .91                    |
| Correlation of one lay rater with one journalist | 0.45            | 0.50                    | 0.50                   |
| Lay raters equivalent to one journalist     | 5.46            | 3.55                    | 4.38                   |
| Lay raters equivalent to two journalists   | >54             | 16.22                   | >54                    |
| Liberal raters equivalent to one journalist | 4.70            | 2.67                    | 2.78                   |
| Liberal raters equivalent to two journalists | >18             | 13.60                   | >18                    |
| Conservative raters equivalent to one journalist | 7.95            | 4.78                    | 6.66                   |
| Conservative raters equivalent to two journalists | >18             | >18                     | >18                    |

Table 10: Summary of results for article collection one only, using the journalist ratings from Allen et al. [2021]. Note that it is only possible to compute equivalences to panels of one or two journalists, because only three journalist ratings were available and one journalist always has to be held out as the reference rater.
|                               | C1   | C2   | C3   |
|-------------------------------|------|------|------|
| Liberal-Conservative Correlation | .83  | .83  | .85  |
| Correlation of one lay rater with one journalist | 0.46 | 0.52 | 0.53 |
| Lay raters equivalent to one journalist | 3.83 | 2.87 | 3.81 |
| Lay raters equivalent to two journalists | 14.22 | 5.42 | 9.66 |
| Lay raters equivalent to three journalists | >54  | 10.30| 40.02|
| Liberal raters equivalent to one journalist | 2.26 | 1.77 | 2.38 |
| Liberal raters equivalent to two journalists | 4.77 | 3.06 | 4.86 |
| Liberal raters equivalent to three journalists | 10.57| 5.03 | 9.29 |
| Conservative raters equivalent to one journalist | >18  | 4.65 | 7.48 |
| Conservative raters equivalent to two journalists | >18  | 13.99| >18  |
| Conservative raters equivalent to three journalists | >18  | >18  | >18  |

Table 11: Summary of results for article collection two only, using our four journalist ratings.
|                                | C1: No Research | C2: Individual Research | C3: Collective Research |
|--------------------------------|-----------------|-------------------------|-------------------------|
| Liberal-Conservative Correlation | .69             | .73                     | .79                     |
| Correlation of one lay rater with one journalist | 0.28            | 0.33                    | 0.33                    |
| Lay raters equivalent to one journalist | 16.49           | 3.86                    | 4.60                    |
| Lay raters equivalent to two journalists | >54             | 7.71                    | 13.86                   |
| Lay raters equivalent to three journalists | >54             | 18.82                   | >54                     |
| Liberal raters equivalent to one journalist | 4.00            | 2.12                    | 2.41                    |
| Liberal raters equivalent to two journalists | 10.83           | 3.69                    | 4.69                    |
| Conservative raters equivalent to one journalist | >18             | 6.57                    | 10.21                   |
| Conservative raters equivalent to two journalists | >18             | >18                     | >18                     |

Table 12: Summary of results for only the 98 articles from collection two that came from non-mainstream sites.
Figure 8: Power curves for the three conditions for only the 98 articles from collection two that came from non-mainstream sites. The x-axis is the number of turkers. The y-axis is the correlation of the mean of $k$ turkers’ ratings with a journalist’s rating. The green horizontal lines show the correlation of a randomly selected journalist’s rating for each item with a held-out journalist (0.47). The blue lines show the correlation of the mean of three journalists with a held-out journalist (0.57).
INFORMED CROWDS CAN EFFECTIVELY IDENTIFY MISINFORMATION

|                                      | C1: No Research | C2: Individual Research | C3: Collective Research |
|--------------------------------------|-----------------|--------------------------|-------------------------|
| Liberal-Conservative Correlation     | .69             | .81                      | .79                     |
| Correlation of one lay rater with one journalist | 0.42            | 0.43                     | 0.42                    |
| Lay raters equivalent to one journalist | 5.77            | 5.19                     | 7.22                    |
| Lay raters equivalent to two journalists | 14.83           | 15.26                    | >54                     |
| Liberal raters equivalent to one journalist | 2.20            | 2.61                     | 2.69                    |
| Liberal raters equivalent to two journalists | 4.69            | 5.31                     | 7.69                    |
| Conservative raters equivalent to one journalist | >18             | >18                      | >18                     |
| Conservative raters equivalent to two journalists | >18             | >18                      | >18                     |

Table 13: Summary of results for only political articles from collection one, using our four journalist ratings.
Figure 9: Power curves for the three conditions for only the 109 political items. The x-axis is the number of turkers. The y-axis is the correlation of the mean of $k$ turkers’ ratings with a journalist’s rating. The green horizontal lines show the correlation of a randomly selected journalist’s rating for each item with a held-out journalist (0.60). The blue lines show the correlation of the mean of three journalists with a held-out journalist (0.69).
INFORMED CROWDS CAN EFFECTIVELY IDENTIFY MISINFORMATION

https://www.lifezette.com/2018/10/kavanaugh-turned-down-scads-of-gofundme-dollars-blasey-ford-hits-paydirt/
https://www.politico.com/states/florida/story/2017/12/15/experts-browards-elections-chief-broke-law-in-destroying-ballots-150258
https://crooksandliars.com/2019/12/devin-nunes-shamelessly-lies-when-hannity
https://www.westernjournal.com/ct/watch-slow-mo-video-catches-angry-acosta-shoving-woman-white-house/
http://nypost.com/2017/09/28/department-of-justice-demands-facebook-account-information-of-anti-trump-activists/
https://www.rsnnews.com/3b8ea8e312074/jackson-man-arrested-for-hacking-a-college-computer-and-returning-all-funds-to-students-since-2010.html
https://crooksandliars.com/2019/12/no-punches-pulled-climate-themed-campaign
https://www.washingtontimes.com/news/2019/nov/12/hillary-clinton-i-would-have-been-a-much-better-mo/
https://townhall.com/tipsheet/leahbarkoukis/2020/01/07/climate-change-turns-out-two-dozen-arrested-for-setting-australias-massive-fire-n2559064?utm_content=buffer1d19c&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer
http://tiphero.com/dont-dress-up-your-chickens/
https://www.foxnews.com/politics/pastors-worship-leaders-pray-for-trump-in-oval-office-amid-impeachment-fight

Table 14: URLs used in experiment.