Investigating Improvements to Crowdsourced Claim Matching

Emma Lurie, John Baldwin, Masha Belyi, Sofia Dewar,
Lucy Li, Rajvardhan Oak, and Daniel Rincón

Info 290 Experiments & Causal Inference
Fall 2019

Abstract

Automated fact-checking approaches often depend on crowdworkers labeling training data, and while the quality of those labels are important to the quality of automated fact-checking systems, little attention has been given to crowdworkers' accuracy at this task. This study focuses on claim matching, the process of matching existing fact-check articles to relevant online documents. We aim 1) to explore the ability of crowdworkers to assess relevance in potential claim matching pairs, 2) to measure whether the quality of instructions affects the accuracy of crowdworkers relevance assessments, and 3) to understand whether additional instructions improve workers' confidence in their ratings. We find that providing examples or visual instructions does not systematically improve the accuracy of workers' ratings or inter-annotator agreement, but it does significantly increase workers' confidence in their ratings.

Introduction

In response to fears about the spread of online misinformation, there has been a rapid growth and investment in fact-checking. However, only a small percentage of people who are exposed to problematic online content are presented with corrective information like fact-checks. In the era of big data and artificial intelligence, a key agenda item for the fact-checking movement has been to develop automated fact-checking systems that leverage techniques like machine learning and natural language processing to limit peoples' exposure to misinformation by having platforms present relevant fact-checks alongside problematic content.

Several systems, including ClaimBuster (Hassan et al. 2017) as well as the system outlined in “Relevant Document Discovery for Fact-checking Articles" (Wang et al. 2018), rely on crowdworkers to match existing fact-check articles to relevant pieces of online content. More formally, claim matching is a process in which *fact-check articles* that do not list the original source of a fact-checked statement (i.e. the claimant) are algorithmically assigned to a piece of the online content produced by a certain publisher, which we refer to as the *document.* This approach has the potential to increase the number of stories that have been fact-checked and simplify the process of displaying fact-checks alongside problematic online content.

However, little is known about how crowdworkers perform on the task of claim matching, yet the misinformation and fact-checking community is eager to incorporate more crowdworkers. This study aims to 1) measure the accuracy of crowdworkers labeling the relevance of existing fact-check articles to additional online document, 2) explore whether a causal link exists between the quality of instructions crowdworkers are provided and the quality of labels, and 3) see whether providing examples correlates to higher confidence levels among crowdworkers.

We design a randomized experiment to answer a descriptive question and two causal questions:

1. **What is the accuracy of crowdworkers’ rating of relevance in claim matching?** Workers were on average able to correctly rate 5 of the 8 claim matched pairs, but the accuracy of workers varied greatly from pair to pair.
2. **Does providing detailed instructions improve the precision and accuracy of raters’ decisions?** Providing detailed instructions improves the accuracy of workers on some pairs but not others, and does not improve inter-rater agreement.
3. **Does providing detailed instructions improve the confidence of raters’ decisions?** Providing detailed instructions significantly increases the confidence of raters’ decisions.

Methods

This section will detail the design of the survey provided to crowdworkers, the experimental conditions selected to deploy on Mechanical Turk, and how we operationalized relevance to determine "ground truth" conceptions of relevance.

Survey Design

Best practices around survey designs for crowdworker fact-checking tasks are not well established, so we will provide an outline for our survey design in the following section with a particular emphasis on the two sets of instructions provided to crowdworkers.

All three versions of our survey (one control, and two treatments) presented crowdworkers with eight claim match pairs of a fact-check claim and a news article headline. All eight pairs related to the topic of U.S. immigration, because this subject is commonly discussed on fact-checking websites. Restricting all pairs rated by workers to a single domain also allows us to avoid the possible confounder of topic affecting perceptions of relevance.

Survey Design for All Conditions (control, Treatment 1, Treatment 2)

We provided the following instructions to all participants:

> MUST READ INSTRUCTIONS: You will now be presented with eight pairs of two sentences. The first sentence of the pair is a headline in an online news article. The second will be a headline of a fact-check article that assesses whether a claim is true or not. Your task is to determine whether the fact-check headline is relevant to the claim in the news article headline.
>
> Please note, you do not have to determine whether the fact-check or article is true or false. You are only responsible for determining if the fact-check is relevant to the news article headline. Do not consult any additional sources to determine the relevance.

Workers were all shown the same eight claim match pairs (fact-check article and news article headline) in a randomized order (randomization done by Qualtrics survey). The fact-check and news article headlines are artificial examples that closely follow the structure of fact-check and news article headlines observed in a corpus of fact-checks and news headlines collected by the first author in a previous study. Workers were also asked to express their confidence in each pair's relevance rating on a 5-point Likert scale from “Extremely confident” to “Not confident at all.” Workers were randomly assigned to treatment or control by the Qualtrics survey software pseudo-random number generator.

Workers were recruited on Amazon Mechanical Turk. Consistent with best practices in other crowdsourced labeling tasks (Budak et al. 2017), workers must have previously completed more than 1,000 HITs, a HIT approval rate of greater than 95%, and reside in the United States. We compensated workers $0.60. We had no further involvement in selecting participants for inclusion and assume the process to be random.

We piloted 12 claims with a small set of crowdworkers (N=30) to identify claims for which crowdworkers had the lowest agreement rates, with the assumption that if a treatment effect did exist for providing instructions, it would be most evident with these examples. All chosen pairs are related to each other (see the Operationalizing Relevance section for the definition of *related* documents). The eight selected claims are in the table below. Half of the pairs are relevant, and half of them are not.

Table 1: The eight article/fact-check pairs that were presented to each participant. Select fact-check titles were selected or modified from the following sources: [*]Snopes, [†]Breitbart. Select article claims were also based on real headlines: [1]press release from Congressman Mark Walker Media Center, [2]Patch Across Texas Health & Fitness, [3]KTLA 5 News, [4]South China Morning Post.

Article Claim	Fact Check Title	Relevance Assessment
The Border is Seeing a Surge of Hepatitis A Cases	A Major Hepatitis A outbreak in San Diego has been pinned on undocumented immigrants there.*	Yes
ICE agents arrest a Tennessee man with multiple past felony convictions	An ICE agent gave up trying to arrest a Tennessee man who, aided by neighbors, refused to leave his vehicle for four hours.*	Yes
House Democrats Just Passed Legislation that Puts Illegal Immigrants Before American Veterans[1]	In September 2019, U.S. House Democrats voted for a bill that would give immigrants on the southern border a better health-record system than veterans.*	Yes
Denver Herald: Immigration Up Since Trump Took Office	The Liberal Daily is right that immigration has decreased since 2016.	Yes
Immigrant Voters in California Support Hillary	An 'illegal immigrant' was convicted of voter fraud for voting multiple times for Hillary Clinton.*	No
Many Texas Cities Susceptible To Large Measles Outbreaks[2]	Immigrants who illegally crossed into the U.S. at the Mexico border are the cause of measles outbreaks in 2019 in the U.S.*	No
U.S. Customs Officer in Texas Loses Job, Citizenship Over His Mexican Birth Certificate[3]	The State of Texas announced it would no longer be issuing birth certificates to the children of undocumented immigrants.*	No
Elizabeth Warren Takes Center Stage as	Elizabeth Warren falsely claims illegal immigration is	No

Economics and Immigration Dominate First Democratic Primary Debate[4]	‘man-made crisis'. †	

Treatment 1: Examples

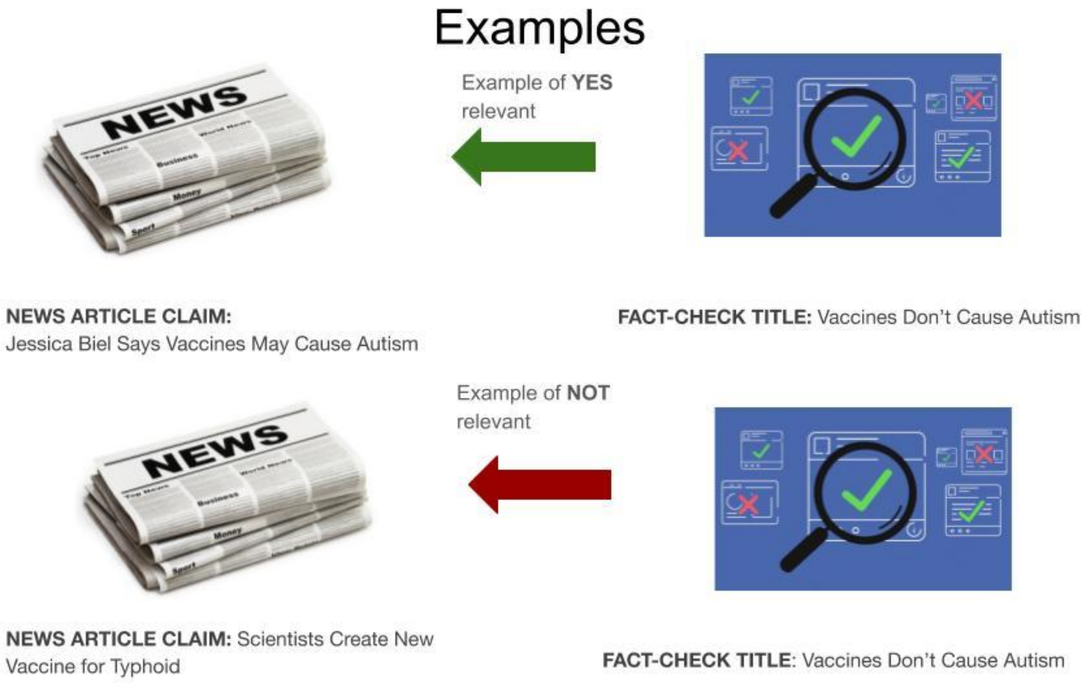
Previous work in psychology has emphasized the importance of supplementing instructions with examples (LeFevre & Dixon 2009; Catrambone 1995). In some cases, participants may believe that examples provide more important information than instructions, and even disregard instructions when examples are present (LeFevre & Dixon 2009). On some crowdsourcing platforms, such as Figure Eight (previously named Crowdflower), workers are tested on example annotations that are already labeled and are notified immediately whether they correctly labeled the instance or not (Le et al. 2010; Sabou et al. 2014). In general, the addition of examples in crowdsourcing tasks is common, and they have been shown to improve worker agreement, especially with language understanding tasks, as well as reduce the amount of time it takes for a task to be chosen by workers (Jain et al. 2017). Thus, we hypothesize that including written examples should also improve accuracy and agreement for the task of determining relevant pairs.

If a crowdworker was randomly assigned to receive Treatment 1, which provided examples of relevant and not relevant pairs, they were shown the following text above the headline pair:

> A relevant fact-check headline does not need to be an exact match, but it does need to "align in spirit" with the news article headline.
>
> Two examples of relevant pairs:
>
> 1. News article claim: "Vaccines May Cause Autism" —> Fact-check: "Vaccines Don’t Cause Autism"
> 2. News article claim: "Jessica Biel Says Vaccines May Cause Autism" —> Fact-check: "Vaccines Don’t Cause Autism"
>
> Two examples of not relevant pairs:
>
> 1. News article claim: "Scientists Create New Vaccine for Typhoid" —> Fact-check: "Vaccines Don’t Cause Autism"
> 2. News article claim: "Vaccines Cause Breast Cancer" —> Fact-check: "Vaccines Don’t Cause Autism'

Treatment 2: Examples + Visualization

Treatment 2 differs from Treatment 1 because it includes a visualization (Figure 1), which depicts the different sources of the two statements that make a pair: news and fact-checkers. We intended for this visualization to clarify the motivation behind obtaining ratings for relevance and as a result, produce relevance ratings that align more closely with the goals of fact-checking. Previous work has shown that the combination of both visual and written information is more effective for learning than written information alone (Mayer 1999), and infographics that include human recognizable objects improve memorability (Borkin et al. 2013).



Examples

Example of **YES** relevant

NEWS ARTICLE CLAIM: Jessica Biel Says Vaccines May Cause Autism

FACT-CHECK TITLE: Vaccines Don't Cause Autism

Example of **NOT** relevant

NEWS ARTICLE CLAIM: Scientists Create New Vaccine for Typhoid

FACT-CHECK TITLE: Vaccines Don't Cause Autism

Figure 1: The visual instructions provided in Treatment 2 of our main experiment for determining relevant and non-relevant fact check and news article pairs.

Operationalizing Relevance

To assess crowdworkers ability to identify relevant claim matches, we must settle on a definition of relevance. As Introna and Nissenbaum (2000) wrote, “determining relevancy is an extraordinarily difficult task...Besides the engineering challenges, experts must struggle with the challenging of approximating a complex human value.”

We model our conception of relevance to be similar to that of Wang et al. (2018)’s definitions:

> Given a fact-checking article with claim *c*, a claim relevant document is a related document that addresses *c*.

Where a related document is defined as:

> Given a fact-checking article, a related document is a document that bears some topical or lexical similarity to the fact-checking article.

Wang et al. (2018) also specifies that “the claim-relevance discovery problem does not require the literal or precise claim to appear in the document but rather aims to find ones that seem to align in spirit.”

A portion of our results depend on our operationalization of relevance. For the eights selected claim matches, three authors had lengthy discussions about whether each claim match pair, given the above definitions, should be rated as relevant or not relevant. In the case of disagreements, a majority vote was taken.

It is precisely because reasonable people disagree about the operationalization of relevance that we hypothesized that providing more robust instructions that include examples may increase crowdworker precision and accuracy when labeling relevance in claim matching.

However, this does not mean that developing a compelling definition of relevance is not important to the success of claim matching systems. When Google’s claim-matching Reviewed Claims system faced heavy scrutiny for applying irrelevant fact-checks to hyperpartisan news stories from *The Daily Caller* and *The Federalist,* Google removed the feature. The official statement from Google explained after finding “that [they] encountered challenges in [their] systems that maps fact checks to publishers, and on further examination it’s clear that we are unable to deliver the quality we’d like for users.”[1] Whether this system faced quality concerns due to a poor operational definition of relevance or other reasons (e.g. poor quality training data, larger cultural problems in the fact-checking ecosystem), it is clear that a successful system must be perceived as accurate and trustworthy.

Alice Marwick, in a larger critique of current fact-checking practices, points to the Reviewed Claims feature as a blunder that may have decreased overall trust in fact-checking (Marwick 2018). Of course, the success of fact-checking depends on readers’ viewing the fact-checks as trustworthy. If fact-checks are not viewed as credible, then they will not influence people’s perceptions of fact-checked content. A priority of the global fact-checking community has been to develop ways to increase readers trust, and good operational definition of relevance is essential to those efforts.

Pilot

We first ran a pilot survey to assess the general difficulty of the relevance matching task, as well as the strength of effect of Treatment 1 (Examples) on relevance matching performance. 50 participants were randomly assigned into control (28) and treatment (22). The median survey completion time was 3.5 minutes.

We found no significant difference in the accuracy of response between the treatment and control groups (Table 2). The average number of correct responses in the control and treatment groups was 5.60 and 5.59 out of 8 respectively.

[1] https://www.poynter.org/fact-checking/2018/google-suspends-fact-checking-feature-over-quality-concerns/

Upon closer inspection, we found that over 90% of the participants in the control group responded correctly to 2 out of the 8 relevance-matching questions, which limited our ability to observe a significant treatment effect. We proceeded to conduct a second control-only pilot asking participants to rate more headline pairs to help us identify more difficult headline pairs to include in our main experiment.

We also discussed blocking by political affiliation as in some previous misinformation-related studies. However, we noticed in the pilot a substantial number of Independents in our sample and decided that it would be more beneficial not to limit the sample to an equal number of Democrats or Republicans through blocking, and rather see the political affiliation breakdown of crowdworkers and add political affiliation as a covariate in our analysis.

Table 2: Pilot Results. There is no significant treatment effect on accuracy.

	Accuracy
Intercept	5.607*** (0.231)
treatment	-0.016 (0.349)
Observations	50
R^2	0.0
Adjusted R^2	-0.021
Residual Std. Error	1.225(df = 48.0)
F Statistic	0.002(df = 1.0; 48.0)

*p<0.1; **p<0.05; ***p<0.01

Power Analysis

Before running the full experiment, we estimated the sample size necessary to observe a significant treatment effect Since our first pilot (N=50, see Pilot Results for further discussion) barely had any treatment effect, we struggled to generate a realistic effect estimate to base our power analysis on. With the assumption that the eight headline pairs in our full experiment would be more challenging, we presumed a treatment effect of 1 with standard deviation=2. In other words, we expected the average number of correct relevance ratings to differ by 1+/-2 between the control and treatment conditions.

We calculated the statistical power our experiment would yield with N ranging from 0 to 400 participants (Figure 2) and found that sample size of 125 would be appropriate for the full experiment with one treatment condition. Coupling this result with cost constraints, we decided to include 225 people in our experiment across two different treatments.

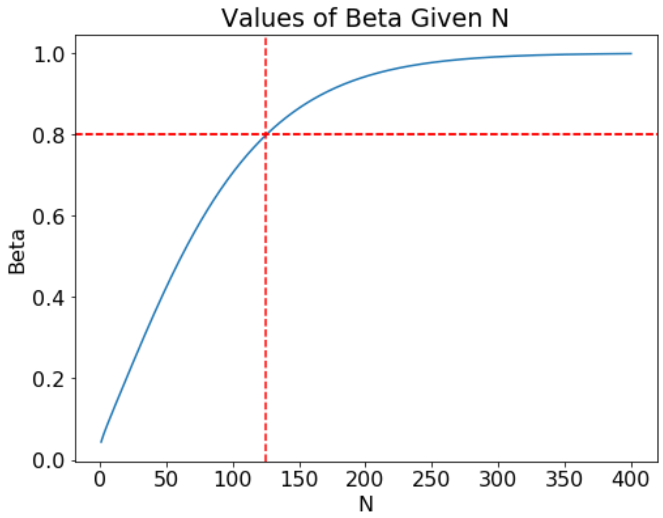


Figure 2: The power curve is the blue line and the red dashed lines represent where the sample size, N, intersects with a β of 80%. We predicted that we would need N=125 where the mean in control was 5 correct claim match labels, the mean in treatment was, and the standard deviation was 2 questions at the 0.05 level.

Experiment

Demographics

From December 7-8, 2019, 204 participants completed the survey with a median completion time of 5 minutes. 260 people started the survey, 56 failed the comprehension check (22%). There was no additional attrition.

- Gender: 126 male (62%), 74 female (36%), 2 genderqueer/non-conforming (1%), 2 prefer to self-describe (1%)
- Education Level: 114 4-year college degree or more (56%), 90 less than a 4-year college degree (44%)
- Political Affiliation: 86 Democrat (41%), 67 Independent (32%), 46 Republican (25%), 3 no preference (1%), 2 other (1%),

Covariate Balance Check

We ran a covariate balance check by regressing our treatment variable against the covariates in our model in an attempt to see if they predicted receiving treatment in any way. We then conducted a F-test between a model with no such covariates and a model with covariates, where a statistically significant difference between these models would indicate covariate imbalance.

The result of our covariate imbalance check showed no signs of covariate imbalance, with the F-test between our models with and without covariates being statistically insignificant (Table 3).

Table 3: ANOVA Comparison of Covariate Models.

Model	F-Statistic	Pr(>F)
Intercept (1)	-	-
1 + Political Affiliation	1.345	0.254
1 + Political Affiliation + Gender	1.642	0.180
1 + Political Affiliation + Gender + Education	1.223	0.291

RQ1: What is the accuracy of crowdworkers rating of relevance in claim matching?

We find that crowdworkers (N=204) were able to correctly rate on average 5 of the 8 claim matched pairs (Table 4). However, as shown in detail in Table 5, despite our efforts in Pilot 2 to identify eight questions that have similar accuracy rates, baseline accuracy ranges from 0.23 to 0.91. Across treatment groups, it seems that some questions are more or less difficult irrespective of survey conditions.

Table 4: Average Number of Correct Claim Matches Labels By Treatment. On average, crowdworkers correctly labeled the relevance of 4.9 claim matches correctly.

	Number of questions correct (standard deviation)
Overall (N=204)	4.97 (1.29)
Control (N=69)	4.72 (1.37)
Treatment 1: Examples (N=70)	5.06 (1.18)
Treatment 2: Visualization + Examples (N=65)	5.14 (1.31)

Table 5: Percent of Crowdworkers Who Correctly Answered Each Question by Treatment. There is a large range of accuracy values across pairs.

	Q1	Q2	pair 3	pair 4	pair 5	pair 6	pair 7	pair 8

All conditions	0.83	0.53	0.88	0.61	0.46	0.28	0.82	0.52
Control	0.91	0.37	0.88	0.52	0.45	0.23	0.84	0.47
Examples	0.84	0.66	0.87	0.68	0.41	0.29	0.82	0.45
Examples + Visualization	0.76	0.56	0.89	0.61	0.51	0.33	0.79	0.64

RQ2: Does providing detailed instructions improve the precision and accuracy of raters' decisions?

Relevance: Accuracy

We find that a simple model containing only the treatment effect variables (Model 0) captures most of the variation of the dependent variable, which is the accuracy of a worker's relevance ratings, as shown by a nested ANOVA test (Table 6). Adding political affiliation (Model 1) does provide some improvement. However, as we can see with the p-values, adding further covariates does not significantly improve the model. No covariates had a significant effect (*p*<0.05) on accuracy, except having "no preference" as political affiliation, but only three workers had that attribute (Table 7).

Table 6: Nested ANOVA Test Results. Model 0 is the base model and contains the treatment variables. Each subsequent model is the previous model augmented with an additional covariate. Model 1 includes political affiliation, Model 2 includes gender, Model 3 includes education level, and Model 4 includes interaction terms of each covariate with treatment. The F statistic value is used to measure whether adding more terms results in a significantly different model.

Model Number	**Model Description**	**F-statistic**	**p-value**
0	**Base model**	NaN	NaN
1	**+ Political affiliation**	2.402437	0.051205
2	**+ PA + gender**	0.079747	0.970938
3	**+ PA + gender + education**	1.049651	0.398074
4	**+ Interaction terms**	0.685484	0.822370

Table 7: Coefficients For Model 0 (Base Model) and Model 1 (Political Affiliation). In Model 0, there seems to be a slight treatment treatment effect for Treatment 2 (p< 0.1). In Model

1, the p-value for "no preference" in political affiliation is significant (*p*<0.05), but the sample size is small (N=3).

	Accuracy	
	Model 0	*Model 1*
Intercept	4.725*** (0.155)	4.848*** (0.225)
Treatment 1: Examples	0.333 (0.218)	
Treatment 2: Examples and Visualization	0.414* (0.223)	
Republican (Treatment 1: Example)		0.276 (0.217)
Republican (Treatment 2: Example and Visualizations)		0.334 (0.222)
Democrat		-0.169 (0.233)
Independent		-0.146 (0.244)
Other		1.347 (0.919)
No Preference		1.837** (0.761)
Observations	204	204
R^2	0.019	0.067
Adjusted R^2	0.009	0.038
Residual Std. Error	1.288(df = 201.0)	1.269(df = 197.0)
F Statistic	1.969(df = 2.0; 201.0)	2.346**(df = 6.0; 197.0)

*p<0.1; **p<0.05; ***p<0.01

We find that there are no significant interaction effects, except for small subgroups. For example, in Model 1 (see Table 7), we find that the "No Preference" political affiliation subgroup has a p-value that is less than 0.05. However, the small size (n=3) leads us to disregard this as a core finding. Hence, overall, we can conclude there is no heterogeneity of treatment effects.

We also find no significant difference between Treatment 1 and Treatment 2: an F-test comparing the two conditions returns F=0.073, *p*=0.787.

We ran the simple model (Model 0) on every pair of headlines separately to see if there were significant treatment effects for some pairs and not others. This analysis is motivated by the fact that there was a lot of variability in the headline wording and relevance types, and as we saw earlier the pairs varied in difficulty. We suspect that our treatment procedures may have helped annotators match only certain types of headline claims and fact-check titles. We observe slightly significant effects of Treatment 1 in pair 2, and significant effects of Treatment 2 in pairs 1, 2, and 8 (Table 8).

Table 8: Treatment Effect per Question. Coefficients range from 0 (wrong) to 1 (correct).

	pair 1	**pair 2**	**pair 3**	**pair 4**	**pair 5**	**pair 6**	**pair 7**	**pair 8**
Treatment 1: Examples	-0.071 (0.061)	0.28* (0.08)	-0.0033 (0.054)	0.16 (0.08)	-0.042 (0.082)	0.061 (0.074)	-0.018 (0.064)	-0.015 (0.081)
Treatment 2: Examples and Visualization	-0.15* (0.061)	0.18* (0.081)	0.0057 (0.054)	0.094 (0.081)	0.061 (0.083)	0.1 (0.075)	-0.054 (0.065)	0.18* (0.082)

* p<0.05

Relevance: Precision (Interrater Agreement)

We find that neither treatment has a positive effect on interrater agreement. As relevance is a complex concept, we hypothesized that including more clarity about the definition of relevance through examples and visualizations would make the crowdworkers more precise. The results of the experiment finds no evidence to support that hypothesis. In fact, the interrater agreement rate is 0.23 (fair agreement) in control and 0.15 (weak agreement) and 0.16 (weak agreement) in Treatment 1 and 2 respectively (see Table 9 for value breakdown).

To measure interrater agreement, we adopt the Fleiss' Kappa statistic, which calculates the degree of agreement in classification compared with that which would be expected by chance. We use the R library raters to calculate interrater agreement. Fleiss' Kappa works well with dichotomous variables and requires more than two raters.

While the Kappa distribution does not have a normal distribution, it is asymptotically normal, therefore, we can use a 2-sample z-test to determine that the agreement rate in treatment is statistically significant from control (p<0.01). However, it is less clear whether there is a practically significant difference in those values. This will be explored further in the discussion.

Table 9: Interrater Agreement by Treatment. Fleiss' Kappa score range from <0 (poor agreement) to 1 (perfect agreement). We find that fair agreement (0.23) in control and weak agreement in both treatment conditions (0.15, and 0.16).

	Fleiss' Kappa
Control	0.23 (0.007)
Treatment 1: Examples	0.15 (0.007)
Treatment 2: Examples and Visualization	0.16 (0.007)

RQ3: Does providing detailed instructions improve the confidence of raters decisions?

Despite a lack of systematic effect on workers' accuracy, both treatments had a much more significant effect on the average confidence per question (Figure 3; Table 11). We performed a nested model analysis to evaluate the effect of treatment and covariates on confidence (Table 10). We found that the most explicative model is the one that only includes treatment. Contrary to the previous model that measured the treatment effects on accuracy, this model has coefficients for both treatment procedures with high statistical and practical significance. The R-squared for this model is also much higher. The effect size is 0.87 and 0.80 respectively for treatment schedules with the examples only and examples+visualizations (measured in confidence units on a 1-5 scale). Though the effect size for treatment with examples only is slightly higher, we found no significant differences between the two treatment types, as supported by comparison with an F-test (F=0.96, p=0.327).

Interestingly, we observed that there is a negative correlation between confidence and accuracy (Figure 4). This correlation is weak and not significant for the aggregated data (pearson coefficient r=-0.099 and p=0.64) and the Control subgroup (pearson coefficient *r*=-0.17 and *p*=0.67) but is much stronger for both Treatment 1 (pearson coefficient *r*=-0.54 and *p*=0.16) and Treatment 2 (pearson coefficient *r*=-0.71 and *p*=0.05).

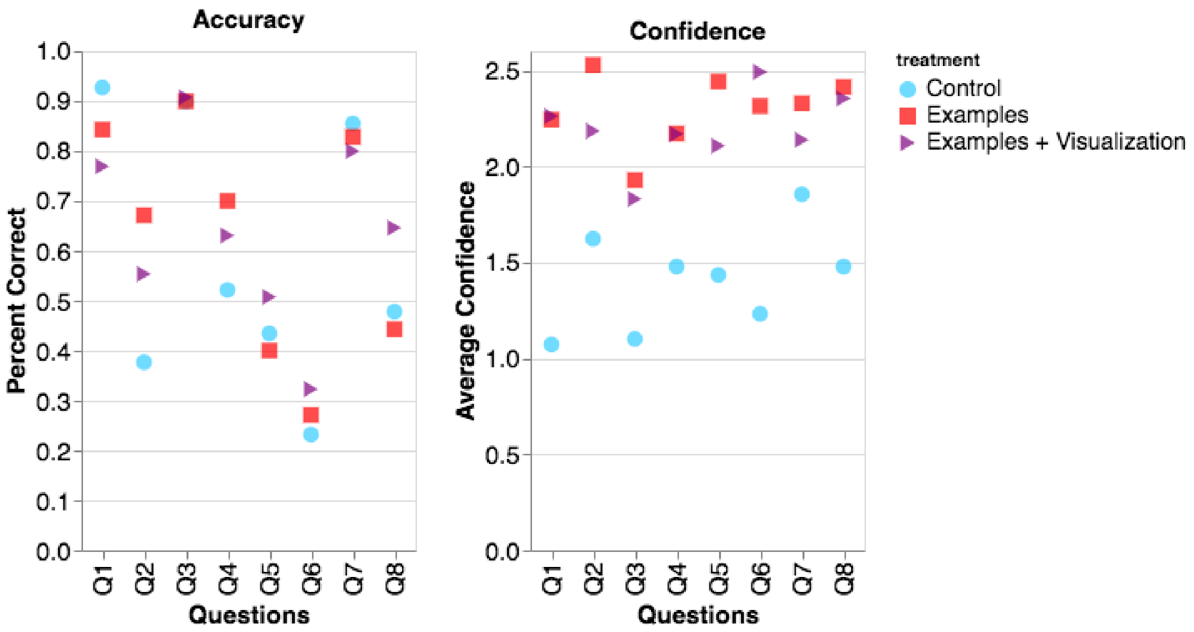


Figure 3: Accuracy of workers per question in the three experimental conditions (left), compared with confidence of workers per question (right). While there is no consistent increase in accuracy with treatment, there is a clear increase in confidence with treatment.

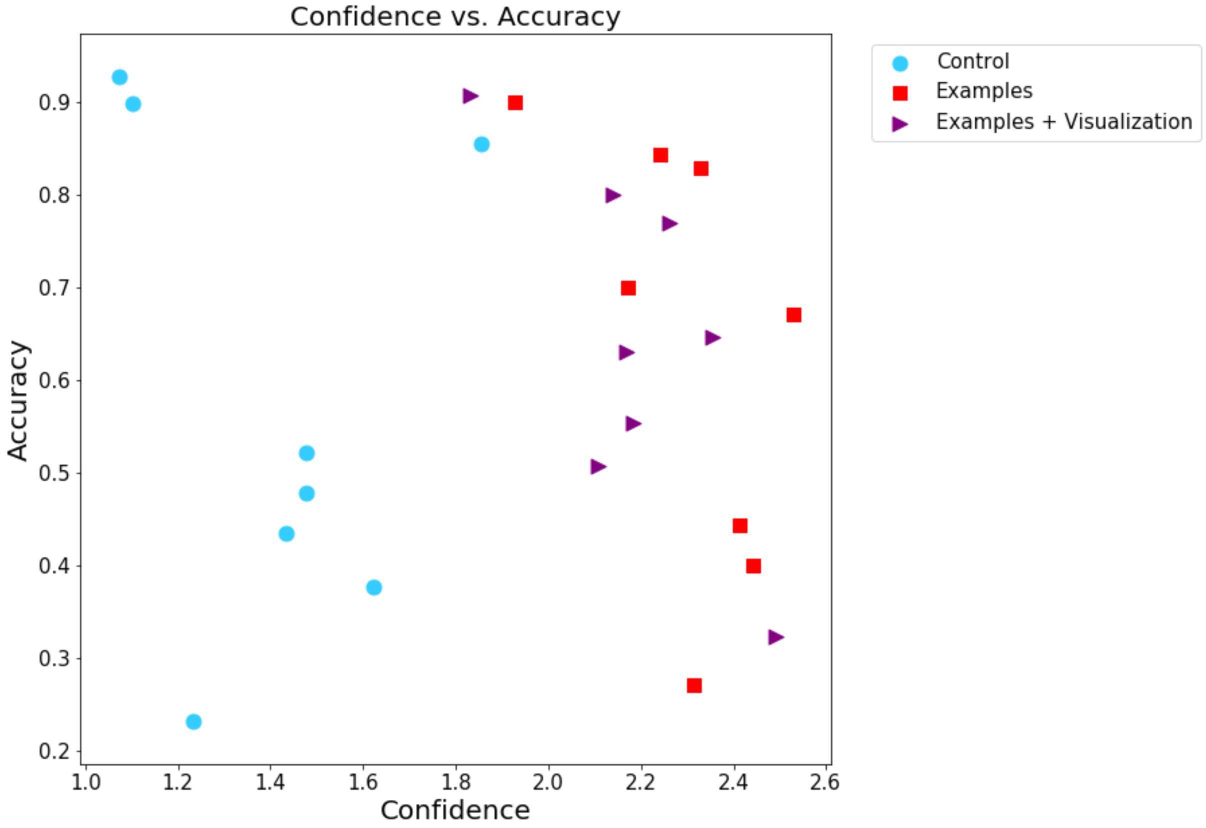


Figure 4: Confidence and accuracy across treatment schedules. There is a slight negative correlation between confidence and accuracy across all conditions, but this correlation significantly more negative in the treatment conditions.

Table 10: Nested ANOVA Test Results. Models 0, 1, 2, 3, and 4 are the same as Table 6 for accuracy, but the dependent variable is instead average confidence. We find that no covariate explains the difference in confidence assessments.

Model #	**Model Description**	**F-statistic**	**p-value**
0	**Base model**	NaN	NaN
1	**+ Political affiliation**	1.848828	0.121024
2	**+ PA + gender**	0.367984	0.776193
3	**+ PA + gender + education**	0.564069	0.784452
4	**+ Interaction terms**	0.883105	0.599588

Table 11: Coefficients for the effect of treatment where the outcome variable is average confidence. We find a highly significant effect of treatment on confidence.

	Confidence
Intercept	1.4094*** (0.059)
Treatment 1: Example	0.8870*** (0.082)
Treatment 2: Example and Visualizations	0.7829*** (0.084)
Observations	204
R^2	0.405
Adjusted R^2	0.399
F Statistic	68.47

*p<0.1; **p<0.05; ***p<0.01

Discussion

With regard to our first research question, we find that crowdworkers were able to successfully identify an average of 5 of 8 claim matches pairs. However, in the control condition, 5 of the 8 claim match pairs were correctly labeled by less than half of crowdworkers (see Table 4). This raises concerns with the status quo method in the automated fact-checking community of taking the majority vote of a handful of crowdworkers rating of whether a claim match pair is relevant. We believe that further research and conversation is needed about appropriate accuracy rates among crowdworkers for labeling relevance in claim matching.

Of course, our accuracy rates are dependent on our operationalization of relevance, and it is reasonable to disagree with our operationalization. We would be interested in rerunning this experiment with different definitions of relevance to relabel our data as well as to present to crowd workers. However, we should note that we attempted to match our definitions of relevance to what has been previously written by others in the fact-checking community about relevance, especially Wang et al.'s (2018) work.

As for our main experiment, which sought to measure the effect of more explicit instructions (including examples and visualizations) on crowdworker performance, we find no statistically significant treatment effect for the treatment with examples (Treatment 1) and a slightly significant treatment effect ($p<0.10$) for the treatment with examples and visualizations (Treatment 2). However, there is no statistically significant difference between the two treatments. Overall, though we find that the effect of treatment is less than half a question's improvement, which seems to have little to no practical significance.

One of our initial intuitions was that rater agreement rates should increase in the treatment condition. As relevance is such a complex concept, providing additional clarification about the concept should make crowdworkers more consistent in their determination. We used a measure of inter annotator agreement, Fleiss' Kappa, shown in Table 9 to compare the effect of treatment in interanotator agreement. We found that crowd workers had fair agreement in the control condition (0.23) and weak agreement in both treatment groups (0.15, 0.16 respectively). One explanation for the decrease in agreement in the treatment condition is that the pairs that had baseline high accuracy rates (and therefore high agreement rates) often had lower accuracy for treatment groups (and therefore slightly lower agreement scores). We found no evidence to support the hypothesis that including examples and visualizations in instructions improves agreement among crowdworkers.

Despite not identifying any increases in crowdworker accuracy as a result of treatment, treatment made crowdworkers significantly more confident (see Figure 3; Table 11).Therefore, it seems that treatment increased confidence without increasing accuracy. In fact, in the treatment conditions, confidence was moderately (Treatment 1) to strongly (Treatment 2) negatively correlated with accuracy. While further research is needed to explore this conclusion, it does match up with previous work that finds that people with some knowledge (compared to total novices) are often overconfident (Sanchez and Dunning, 2018). As automated fact-checking systems continue to experiment with the ideal set up for similar tasks, this is a finding that warrants further attention.

Limitations

Our study has several limitations. One is that the questions are limited to the immigration domain. While we selected immigration because it's a topic that 1) appears in fact-checks and 2) is likely to be familiar to crowdworkers, it is also not clear how generalizable the lessons from this study are to health, climate, or political misinformation more broadly.

Additionally, while we attempted to either use or closely imitate existing fact-check articles and news articles, some of the examples are artificial. Moreover, there are only eight examples to try to minimize worker fatigue while still maintaining a large number of crowdworkers viewing each pair, but the limited number of examples may mean that headline wording affects our results.

Another potential limitation of the study is that the various questions have different baseline accuracy rates. For questions that have high accuracy rates (around 0.90), it seems that treatment decrease accuracy. However, there is not much room for treatment to have a positive impact with high baseline examples. Adding more questions with lower baseline accuracy rates, may change our results. It is also possible that the example pairs we provide to workers are not comparable in difficulty to the pairs workers are asked to rate in live claim matching systems.

As previously mentioned, many of our conclusions rely on accuracy, which is scored based on our definition of relevance. Other definitions may find different results.

Conclusion

As we build reliance on human-in-the-loop fact-checking ML systems, it is important to test the robustness and assumptions of the crowdsourcing tasks we design. The definition of *relevance* in the context of fact-checking is a murky one and the language of news article headlines and fact check claims can be related in complex ways. Our primary goal was to investigate whether providing instructions to workers would improve their performance in the claim-matching task. We found that providing examples of relevant and non-relevant pairs does not have a systemically significant effect on workers' accuracy or precision, but does increase their confidence. These results show that there is more to the variance in workers' performance than can be explained by the presence of examples. Given the wide variance we see across questions, future work should identify how different subtypes of relevance might affect the outcome variables we studied here. Further studies that methodically change different linguistic aspects of pairs may shed more insight into the weaknesses and strengths of crowdsourcing claim matching.

References

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2306-2315.

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), 250-271.

Catrambone, R. (1995). Following instructions: Effects of principles and examples. *Journal of Experimental Psychology: Applied*, *1*(3), 227.

Hassan, Naeemul, et al. "Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

Introna, Lucas, and Helen Nissenbaum. "Defining the web: The politics of search engines." *Computer* 33.1 (2000): 54-62.

Jain, A., Sarma, A. D., Parameswaran, A., & Widom, J. (2017). Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment*, *10*(7), 829-840.

Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010, July). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation* (Vol. 2126, pp. 22-32).

LeFevre, J. A., & Dixon, P. (1986). Do written instructions need examples?. *Cognition and Instruction*, *3*(1), 1-30.

Mayer, R. E. (1999). Multimedia aids to problem-solving transfer. *International Journal of Educational Research*, *31*(7), 611-623.

Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014, May). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (pp. 859-866).

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning dangerous thing?. *Journal of Personality and Social Psychology*, *114*(1), 10.

Wang, Xuezhi, et al. "Relevant document discovery for fact-checking articles." *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018.