# Investigating Causal Effects of Instructions in Crowdsourced Claim Matching

**Emma Lurie**
Univeristy of California, Berkeley
emma_lurie@berkeley.edu

**Li Lucy**
Univeristy of California, Berkeley

**Masha Belyi**
Univeristy of California, Berkeley

**Sofia Dewar**
Univeristy of California, Berkeley

**Daniel Rincón**
Univeristy of California, Berkeley

**John Baldwin**
Univeristy of California, Berkeley

**Rajvardhan Oak**
Univeristy of California, Berkeley

## ABSTRACT

Increasing the reach of fact-check articles is a common strategy to better combat online misinformation. This study focuses on claim matching, the process of matching existing fact-check articles to relevant online documents, which can be classified as a type of automated fact-checking. Automated fact-checking approaches, including claim matching, 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. We aim 1) to explore the ability of crowdworkers to assess relevance in potential claim matching pairs, 2) to measure whether the length and quality of instructions affects the accuracy of workers' relevance assessments, and 3) to understand whether additional instructions improve workers' confidence in their ratings. We find that providing written or visual instructions does not improve the accuracy of workers' ratings or inter-annotator agreement, but it does significantly increase workers' confidence in their ratings.

## 1 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.

*C+J '20, March 2020, Boston, MA, USA*


Several systems, including ClaimBuster [4] as well as the system outlined in "Relevant Document Discovery for Fact-checking Articles" [11], rely on crowdworkers to match existing fact-check articles to relevant pieces of online content. Formally, claim matching is a process in which a fact-check article is algorithmically assigned to a piece of relevant online content. This approach has the potential to increase the amount of content that have been fact-checked and simplify the process of displaying fact-checks alongside problematic online content.

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

We design a randomized experiment (N=204) to answer a descriptive question as well as a causal question:

(1) **What is the accuracy of crowdworkers' rating of relevance in claim matching?** Workers were on average able to correctly rate five of the eight claim matched pairs, but the accuracy of workers varied greatly from question to question.
(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.

## 2 OPERATIONALIZING RELEVANCE

To assess crowdworkers ability to identify relevant claim matches, we must settle on a definition of relevance. As Introna and Nissenbaum [5] write, "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.'s [11] 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. [11] also specify 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."

For the eight 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 hypothesize that providing more robust instructions that include examples may increase crowdworker precision and accuracy when labeling relevance in claim matching.

## 3 SURVEY DESIGN

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 (Table 1) related to the topic of U.S. immigration. We selected immigration because this subject is a common topic on fact-checking websites as well as in the public debate. Restricting all pairs rated by workers to a single domain allows us to avoid the possible confounder of topic affecting perceptions of relevance.

Crowdworkers were all shown the same set of standard instructions which explained that the task was to assess the relevance of the claim match pair rather than the accuracy of specific content. Crowdworkers were then given two question comprehension task to ensure that they understood the task. If they passed the comprehension task, they were able to proceed assessing the eight claim match pairs.

We recruited workers on Amazon Mechanical Turk. Consistent with best practices in other crowdsourced labeling tasks (e.g. [2]), 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.

Workers were all shown the same eight claim match pairs (fact-check article and news article headline) in a randomized order. Workers were also asked to express their confidence in each pairs' relevance rating on a 5-point Likert scale from "Extremely confident" to "Not confident at all." The fact-check and news article headlines were a combination of actual claims/headlines and artificial examples that closely follow the structure of fact-check and news article headlines observed in a corpus collected by the first author in a previous study.

### Treatment 1: Examples

Previous work in psychology has emphasized the importance of supplementing instructions with examples [3, 7]. In some cases, participants may believe that examples provide more important information than instructions, and even disregard instructions when examples are present [7]. 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 [7, 9]. 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 [6]. Thus, we hypothesize that including written examples should also improve accuracy and agreement for the task of determining relevant pairs. The participants assigned to Treatment 1 were provided the instructions shown in Figure 1.

### Treatment 2: Examples + Visualization

Previous work has shown that the combination of both visual and written information is more effective for learning than written information alone [8], and infographics that include human recognizable objects improve memorability [1]. We therefore decided to add a visual component to our treatment. Treatment 2 differs from Treatment 1 in that an infographic containing an example is shown to participants (Figure 2).

## 4 RESULTS

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%).

| Article Claim | Fact Check Title | Relevant |
|---|---|---|
| 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 | 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 | 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 | 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 Economics and Immigration Dominate First Democratic Primary Debate | Elizabeth Warren falsely claims illegal immigration is 'man-made crisis'. | No |

**Table 1: The eight article/fact-check pairs that were presented to each participant.**


**Figure 1: Instructions shown in Treatment 1.**

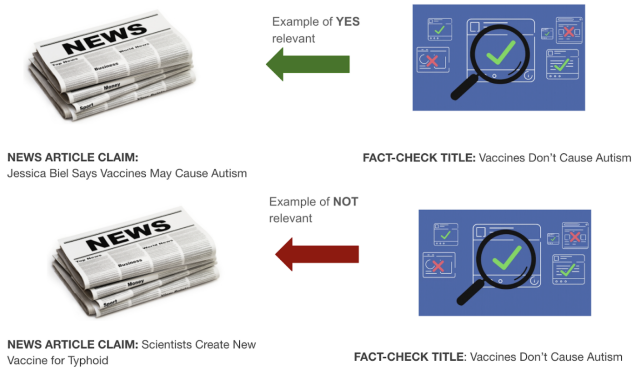


**Figure 2: Infographic added in Treatment 2.**

| | **Number of questions correct** (standard deviation) |
|---|---|
| All conditions (N=204) | 4.97 (1.29) |
| Control (N=69) | 4.72 (1.37) |
| Treatment 1: (N=70) | 5.06 (1.18) |
| Treatment 2: (N=65) | 5.14 (1.31) |

**Table 2: On average, crowdworkers correctly labeled the relevance of 5 claim matches correctly.**

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

Crowdworkers (N=204) were able to correctly rate on average 5 of the 8 claim matched pairs (Table 2). However, baseline accuracy ranges from 0.23 to 0.91 across questions and treatment groups (Table 3). It seems that some questions are more or less difficult irrespective of survey conditions.

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

*Relevance: Accuracy.* Treatment 2 has a slightly statistically significant effect ($p < 0.1$) on accuracy. We perform a nested

| | Q1 | Q2 | Q3 | Q4 | Q4 | Q6 | Q7 | Q8 |
|---|---|---|---|---|---|---|---|---|
| All | 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 |
| Treatment 1 | 0.84 | 0.66 | 0.87 | 0.68 | 0.41 | 0.29 | 0.82 | 0.45 |
| Treatment 2 | 0.76 | 0.56 | 0.89 | 0.61 | 0.51 | 0.33 | 0.79 | 0.64 |

**Table 3: The percent of crowdworkers who correctly answered each question, by treatment. There is a large range of accuracy values across pairs.**

ANOVA test to see whether including covariates (a worker's political affiliation, gender, education level) produce significant models. No F-statistic has a $p$-value of less than 0.05. Thus, we report the ordinary least squares (OLS) regression results of the effect of treatment on accuracy (Table 4).

| | *Accuracy* |
|---|---|
| Treatment 1 | 0.333 |
| | (0.218) |
| Treatment 2 | 0.414* |
| | (0.223) |
| Intercept | 4.725*** |
| | (0.155) |
| Observations | 204.0 |
| $R^2$ | 0.019 |
| Adjusted $R^2$ | 0.009 |
| Residual Std. Error | 1.288 (df = 201.0) |
| F Statistic | 1.969 (df = 2.0; 201.0) |

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

**Table 4: OLS regression results for the effect of treatment on accuracy.**

*Relevance: Precision.* We find that neither treatment has a positive effect on interrater agreement. In fact, the Fleiss' kappa, or agreement rate across workers, is $\kappa = 0.23$ (fair agreement) in control and $\kappa = 0.15$ (weak agreement) and $\kappa = 0.16$ (weak agreement) in Treatment 1 and 2 respectively.

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

Despite a lack of treatment effect on workers' accuracy, both treatments had a significant effect on the workers' average confidence per question (Figure 3). A nested ANOVA to evaluate the effect of treatment and covariates on confidence showed that the most explicative model is the base model that measures the effect of treatment on confidence. This model has coefficients for both treatment procedures with high statistical and practical significance (Table 5). The effect size is 0.87 and 0.80 respectively for Treatment 1 and Treatment 2 (measured in confidence units on a 1-5 scale). Though the effect size for Treatment 1 is slightly higher, we found no significant differences between the two treatment types, as supported by comparison with an F-test ($F = 0.96, p = 0.33$).

The correlation between confidence and accuracy is weak and not significant for the aggregated data ($r = -0.099$ and $p = 0.64$) and the Control subgroup ($r = -0.17$ and $p = 0.67$) but is much stronger for both Treatment 1 ($r = -0.54$ and $p = 0.16$) and Treatment 2 ($r = -0.71$ and $p = 0.05$).

## 5 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. We believe that further research and conversation is needed about appropriate accuracy rates among crowdworkers for labeling relevance in claim matching.

In the control condition, only 3 of the 8 claim match pairs were correctly labeled by the majority of crowdworkers (Table 2). This raises concerns with the status quo method in the automated fact-checking community of taking the majority vote of a handful of crowdworkers ratings to determine whether a claim match pair is relevant.

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 [11] work.

As for our the effect of more explicit instructions (including examples and visualizations), we find no statistically significant treatment effect for Treatment 1 and a slightly significant treatment effect ($p < 0.10$) for Treatment 2. However, there is no statistically significant difference between the two treatments. Overall, 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, but we find no evidence to support this hypothesis. In fact, we found that crowd workers had fair agreement in the control condition ($\kappa = 0.23$) and weak agreement in both treatment groups ($\kappa = 0.15, 0.16$ respectively). This reflects the continued challenge it is to define relevance, and brings into question whether the current methods of crowdworker labeled relevance assessments is providing the desired outcomes.

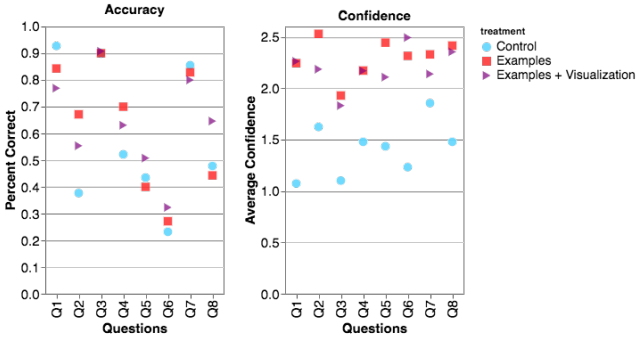


**Figure 3: While there is no obvious interpretation between treatment and accuracy, there is a clear relationship between treatment and confidence level across questions.**

| | *Worker Confidence* |
|---|---|
| Treatment 1 | 0.887*** |
| | (0.082) |
| Treatment 2 | 0.783*** |
| | (0.084) |
| Intercept | 1.409*** |
| | (0.059) |
| Observations | 204.0 |
| $R^2$ | 0.405 |
| Adjusted $R^2$ | 0.399 |
| Residual Std. Error | 0.486 (df = 201.0) |
| F Statistic | 68.475*** (df = 2.0; 201.0) |

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

**Table 5: OLS regression results for effect of treatment on confidence levels.**

Additionally, an important finding of this work is that treatment made crowdworkers significantly more confident (Figure 3; Table 5). Therefore, it seems that treatment increased confidence without increasing accuracy. In fact, in the treatment conditions, confidence level is negatively correlated with accuracy.

While confidence levels increased by an average of almost a 1 point on the 5-point Likert scale in the treatment condition, the average confidence rating in treatment was between 2-2.5, which is still relatively unconfident. Further research is needed to explore this conclusion, but it does match up with previous work that finds that people with some knowledge (compared to total novices) are often overconfident [10]. As automated fact-checking systems continue to experiment with the ideal set up for similar tasks, this is a finding that warrants further attention.

## 6 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. 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 suggest that there is more to the variance in workers' performance than can be explained by the presence of examples. Given the wide variance 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 crowdsourced claim matching.

## REFERENCES

[1] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.

[2] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.

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

[4] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1945–1948.

[5] Lucas Introna and Helen Nissenbaum. 2000. Defining the web: The politics of search engines. *Computer* 33, 1 (2000), 54–62.

[6] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment* 10, 7 (2017), 829–840.

[7] Jo-Anne LeFevre and Peter Dixon. 1986. Do written instructions need examples? *Cognition and Instruction* 3, 1 (1986), 1–30.

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

[9] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. 859–866.

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

[11] Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant document discovery for fact-checking articles. In *Companion Proceedings of the The Web Conference 2018*. 525–533.