# Red Flagging Fake News

Suhas Gupta, Kevin Drever, Imran Manji

## Metrics

For measuring the affect of our treatment we scored the participants on a binary choice of "Yes" or "No" about whether they believed the content of each post displayed during the suvery. Subjects scored one point if they answered correctly ("Yes" for true tweets and "No" for false tweets) and zero points otherwise. We added up the points for all the tweet related questions for each participants and generated the *total score* (score on all tweet questions), *true tweet score* (score on true tweets only) and *false tweet score* (score on false tweets only). These scores were used in regressions to analyze the affect of treatment in this experiment.

## Results

We analyzed two major effecs of our experiment using regression with robust standard errors. First, we tested our primary NULL hypothesis that displaying general warning flags with social media posts doesn't affect people's belief in **false** posts, against the alternate hypothesis that displaying generals warning flags reduces the believability in fake stories. In order to test this, we regressed *true_tweet_score* over the *warning flag* (treatment) to calcuate the treatment effect coefficient and standard errors. Robust standard errors were used throughout the regression tables in this experiment's data analysis. We pooled data from all three surveys to run the regerssions and included indicator variables for subjects from Amazon's Mechanical Turk and whether the survey included a Captcha verification at the beginning. We also ran regressions with question fixed effects (for each tweet question) and interaction terms with Age, Party and Gender that we hypothesized to likely have the most effect on fake news believability. Secondly, we tested the spillover effect of warning flags on true social media posts. Since our treatment only involves displaying a general fake news warning flag, we believed that there might be a reduction in participants' belief in true posts as well since both true and false tweets are displayed in the same survey. We test this effect using the same regression strategy as the primary effect.

### Effect of general warning on false tweets

Figure 1 shows the total mean scores of all 313 survey takers for both treatment (flag) and control (no flag) conditions. Note, that this is the mean of total score, i.e. including score and true and false tweets. The figure indicates a decrease in participant score when the treatment warning flag is present (from 7.16 to 7.05 out of a scale of 10).

Figure 1. Average score for all tweets by survey assignment group

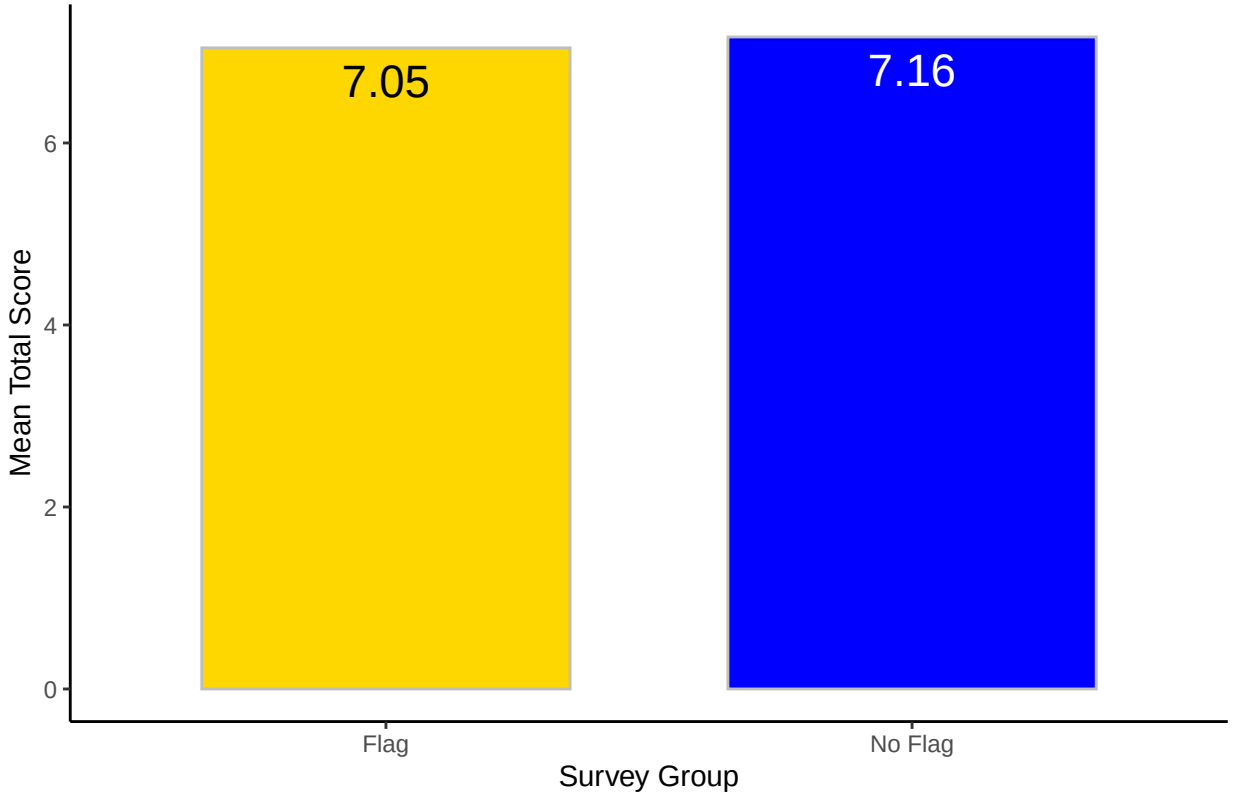
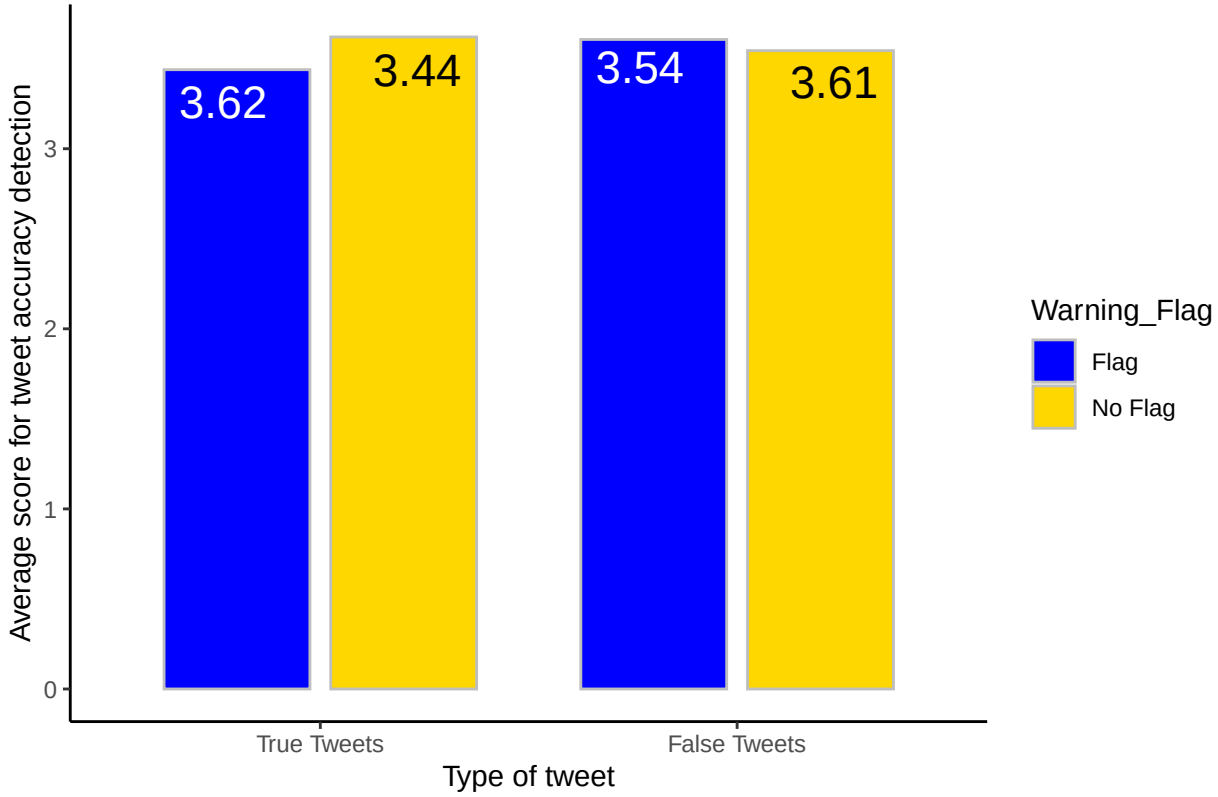



Figure 2. beraks down the point sample mean scores on false and true tweets in the presence and absence of warning flags. We see from this figure that for false tweets, the presence of a flag increases the participants' ability to correctly identify them by 1.97%. However, we also see that the ability to correctly identify true tweets reduces in the presence of the warning flag by 5% as compared to no flag survey. We tested the hypothesis formally in Table 1, which shows pooled regression results (over all 313 participants across 3 surveys) for score on false tweet questions. There appears to be an increase of 6% in score of participants for correctly identifying the false tweets with a confidence interval of (-0.3,0.420) before including control for mechanical turk participants and BOT checks. We added indicator variables for participants recruited on mechanical turk as participants on amazon's mechanical turk may not be accurate representatives of general US population. Futhermore, we also added an indicator variable for a BOT check being present in the survey to control for any malignant activity on mechanical turk (using BOTs to answer survey and earn monetray rewards). Since the CAPTCHA verification was added in the latter half of the experiment, we added an indicator variable in regression to control for errors due to BOT activity. We see that the 95% confidence intervals shrink slightly when we control for these covariates. The results do not show a statistically significant result for the treatment effect [0.061 (0.140)]. This result supports our NULL hypothesis that a general warning flags doesn not have any effect on human ability to detect false posts on twitter.However, consistent with our intuition, the coefficient sign indicates that there is a positive effect on score for detecting false tweets when a warning flag is present.

```
##    Warning_Flag   Score_Type Score
## 1:      No Flag  True Tweets   3.6
## 2:         Flag  True Tweets   3.4
## 3:      No Flag False Tweets   3.5
## 4:         Flag False Tweets   3.6
```

Figure 2. General warning effect on true and false tweets




% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:04

Figure 3 shows that distribution of gender among our survey participants while Figure 4 shows breaks the party affiliation by gender among our survey respondents. We see a fairly balanced distributed between Male and Female survey respondents. The distribution of male and female respondents by party affiliation (Figure 4) also appears to be well balanced for all three party affiliations. We find however, that the distribution is skewed towards the number of democratic party affiliation compared to Republicans or *Other*. Figure 5 indicates that the mean score for detecting the false tweets in the survey is slightly better for males versus females. Figure 6, shows that the mean scores for detecting false tweets correctly is substantially for *Other* party affiliates followed by Democrats and Republicans in order.

Table 1: Table 1. Treatment effect of warning flag on false tweet score

| | Outcome Variable | | |
|---|---|---|---|
| | Score on False Tweets | | |
| | ATE(no indicator) | ATE(Mturk indicator) | Baseline model |
| | (1) | (2) | (3) |
| Warning Flag | 0.062 | 0.056 | 0.061 |
| | (0.180) | (0.170) | (0.140) |
| MTurk Participant | | $-1.200^{***}$ | $-0.180$ |
| | | (0.150) | (0.120) |
| Captcha Verified | | | $2.300^{***}$ |
| | | | (0.190) |
| Constant | $3.500^{***}$ | $4.400^{***}$ | $2.100^{***}$ |
| | (0.130) | (0.130) | (0.220) |
| Question Fixed Effects | No | No | No |
| Observations | 313 | 313 | 313 |
| $R^2$ | 0.0004 | 0.096 | 0.450 |
| Adjusted $R^2$ | $-0.003$ | 0.090 | 0.440 |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Regression models with robust standard errors.
Warning Flag is the treatment and outcome variable is participant on false tweet questions in su
Respondents in treatment saw the flag and those in control didn't see any warning.
Baseline mode is the third column in table with indicator variables for Mturk and BOT verificati

Figure 3. Gender distribution of survey takers

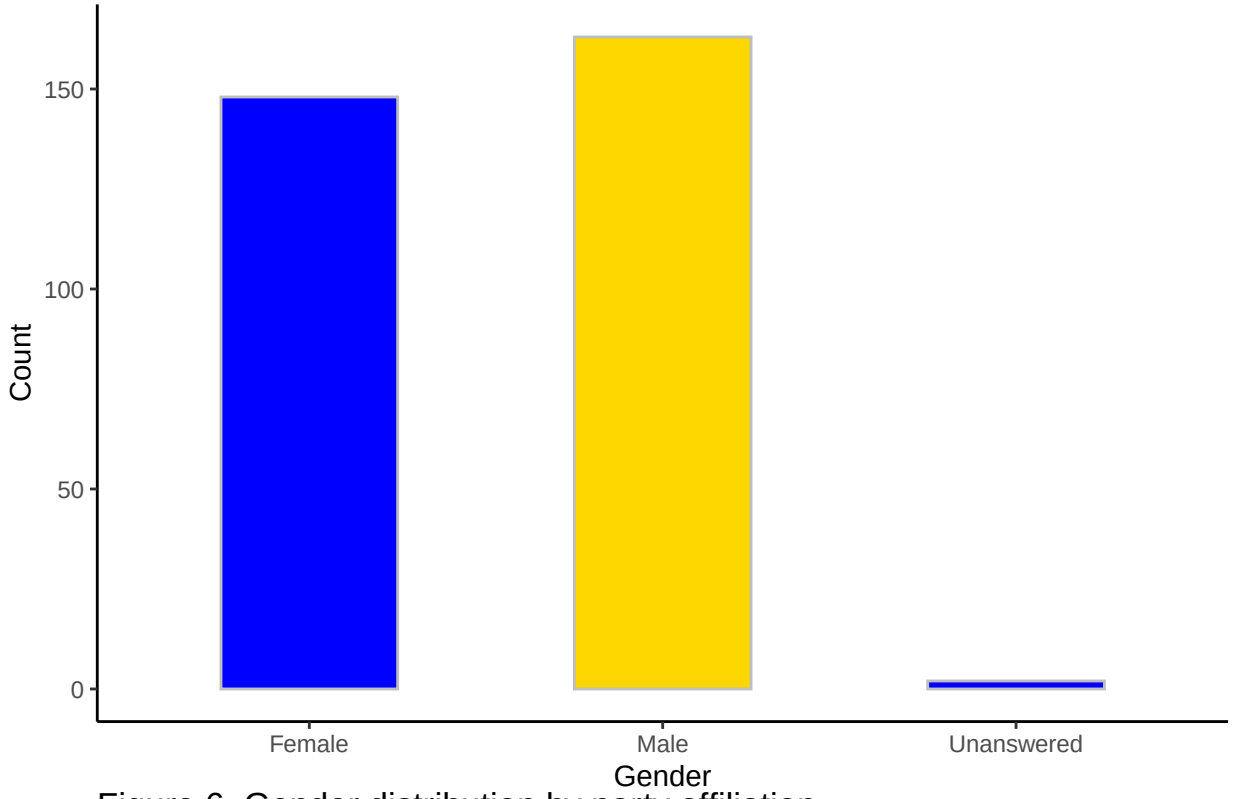



Figure 6. Gender distribution by party affiliation

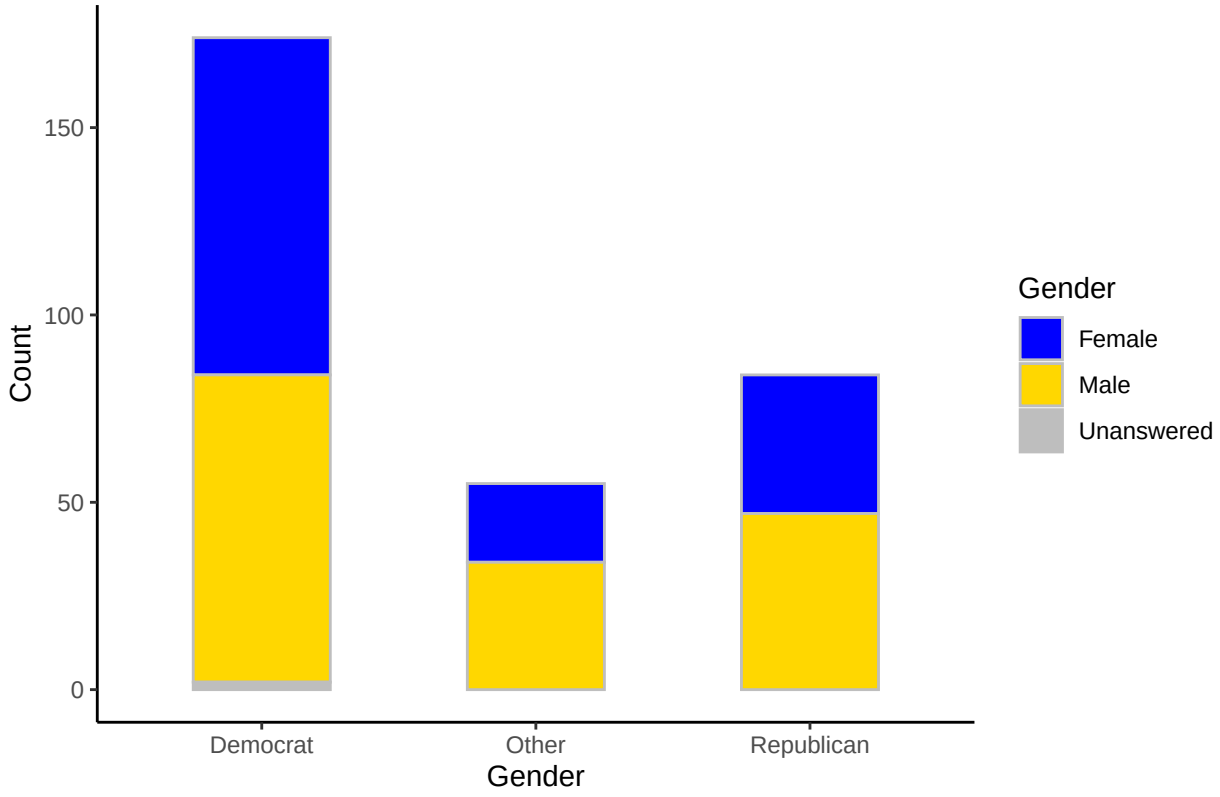


Figure 5. Mean scores by Gender

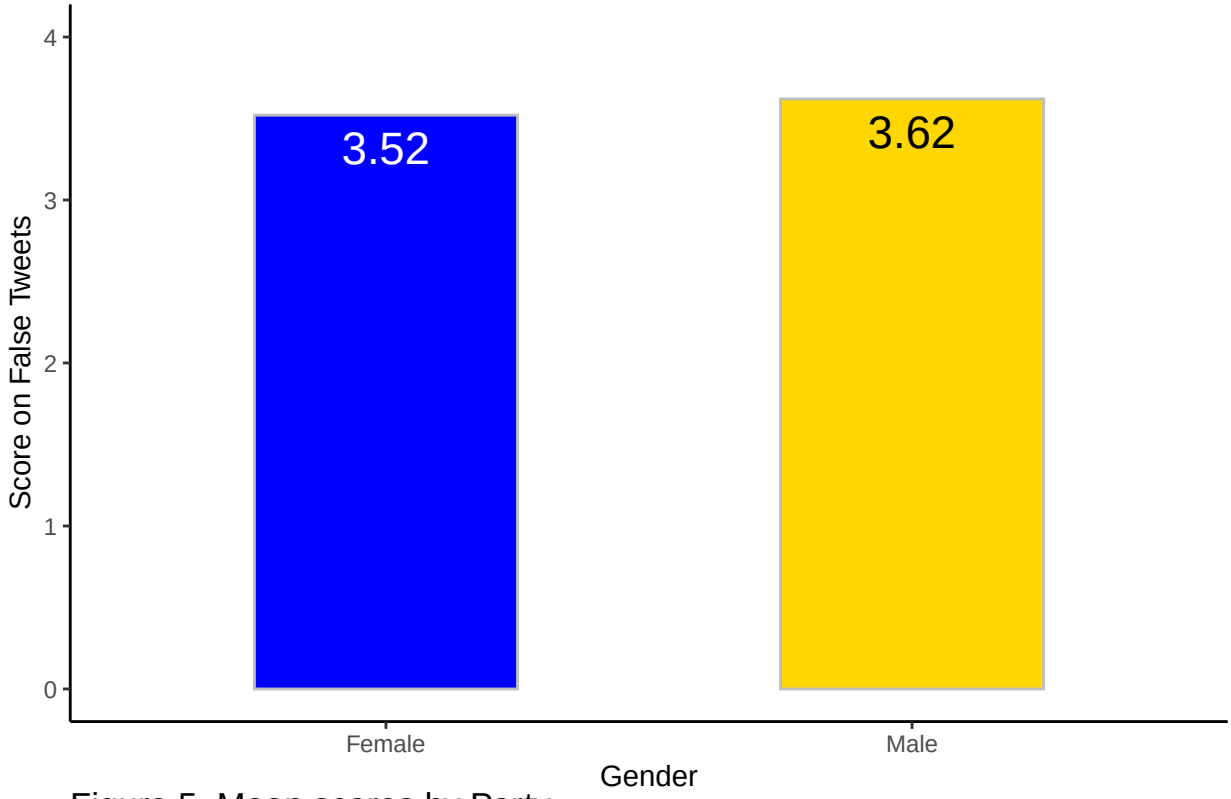



Figure 5. Mean scores by Party

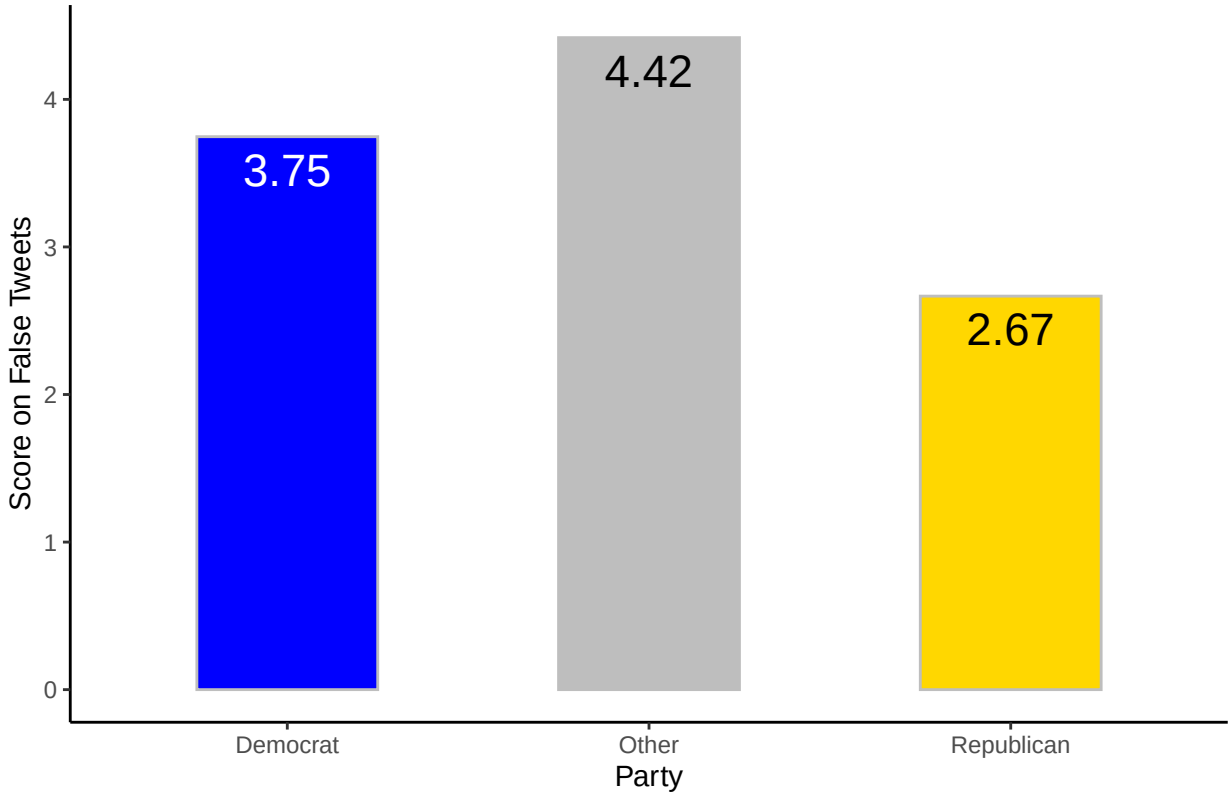
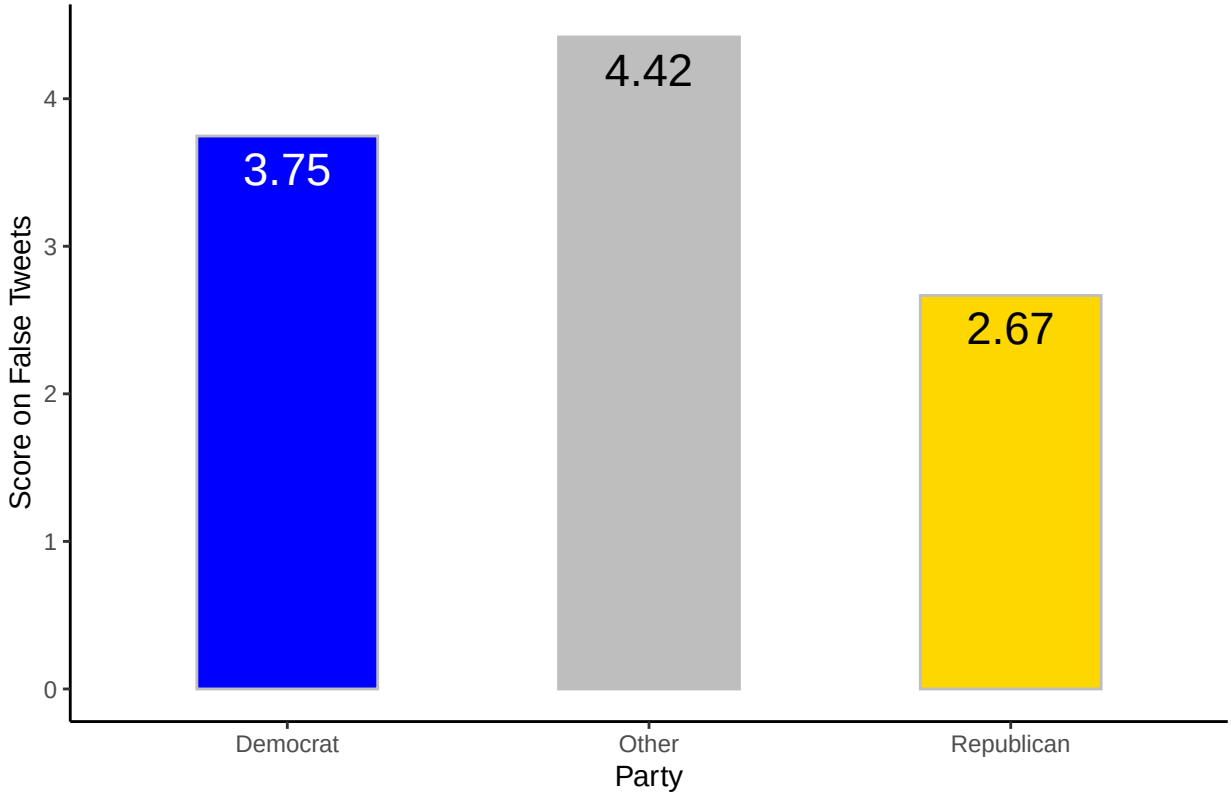


Figure 5. Mean scores by Party




```
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
```

Based on the above point sample data analysis, we are motivated to test the interaction effects gender, party affiliation and age on our outcome variable (score on false tweet detection). Some research has shown that false (especially negative) news coverage may be targeted towards female viewers more than men and there is tendency to believe in different kind of false news based on a person's gender [1]. In our survey, we have attempted to include tweets from political, scienctific and general US recent affairs which makes us hypothesize that the warning flag for false tweets may have different marginal effects on subjects aligning with different political parties. Table 2 summarizes the regression result with interaction terms with gender and party and shows that there is not statistically significant marginal effect of showing the warning flag on accurate detection of false twitter posts.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:08

Next we analyze the age and education co-variates collected during the experiment. It has been shown that older people (65+) tend to share more fake news on social media than people younger in age [2].There is also research indicating that people with higher education are less affected by fake news on social media [3]. Figure 7 shows the distribution of different age groups in our experiment's data set. We have a large representation of respondents within the 21-40 age group followed by the 41-60 group. The 61 and older as well as less than 20 years age are not well represented in our survey results. This skew will reduce the power of the experiment while trying to determine the marginal effects of treatment within the different age groups. Figure 8 reveals that the data set has a high representation of subjets with college education and graduate degrees while lower education categories are not well represented. This is indicative of the fact that the recruits for the experiment were from experimenters personal and professional network (who are

Table 2: Table 2. Treatment effect of warning flag on false tweet score with Gender and Party interaction terms

| | Outcome Variable | | |
|---|---|---|---|
| | Score on False Tweets | | |
| | Marginal effects of Gender | Marginal effects of Party | Gender and Party |
| | (1) | (2) | ( |
| Warning Flag | −0.038 | −0.037 | −0 |
| | (0.190) | (0.170) | (0. |
| MTurk Participant | −1.100*** | | −0.9 |
| | (0.290) | | (0. |
| Captcha Verified | −0.670*** | | −0.7 |
| | (0.220) | | (0. |
| PartyOther | | 0.280 | 0. |
| | | (0.230) | (0. |
| PartyRepublican | | −0.540** | −0 |
| | | (0.260) | (0. |
| Mturk | −0.180 | −0.120 | −0 |
| | (0.120) | (0.120) | (0. |
| captcha | 2.300*** | 2.100*** | 2.1 |
| | (0.190) | (0.190) | (0. |
| assignment:GenderFemale | 0.180 | | 0. |
| | (0.280) | | (0. |
| assignment:GenderMale | | | |
| assignment:PartyOther | | 0.250 | 0. |
| | | (0.300) | (0. |
| assignment:PartyRepublican | | 0.120 | −0 |
| | | (0.350) | (0. |
| GenderFemale:PartyOther | | | 0. |
| | | | (0. |
| GenderMale:PartyOther | | | |
| GenderFemale:PartyRepublican | | | −0 |
| | | | (0. |
| GenderMale:PartyRepublican | | | |
| assignment:GenderFemale:PartyOther | | | −0 |
| | | | (0. |
| assignment:GenderMale:PartyOther | | | |
| assignment:GenderFemale:PartyRepublican | | | 0. |

likely have graduate degress) and from mechanical turk based only in the United States (who have mostly a minimum college education). Figure 10 shows that the mean score is highest for respondents in the age 0-20 group and there is not a large difference in the other age groups. Figure 11 shows that for Mechanical Turk participants, the score is higher for the education category of *some college* while for non Mechanical Turk respondents the highest score was achieved by people with a graduate degree. This in line with the distribution of respondents between the two types of survey respondents.

Figure 7. Age distribution of survey takers

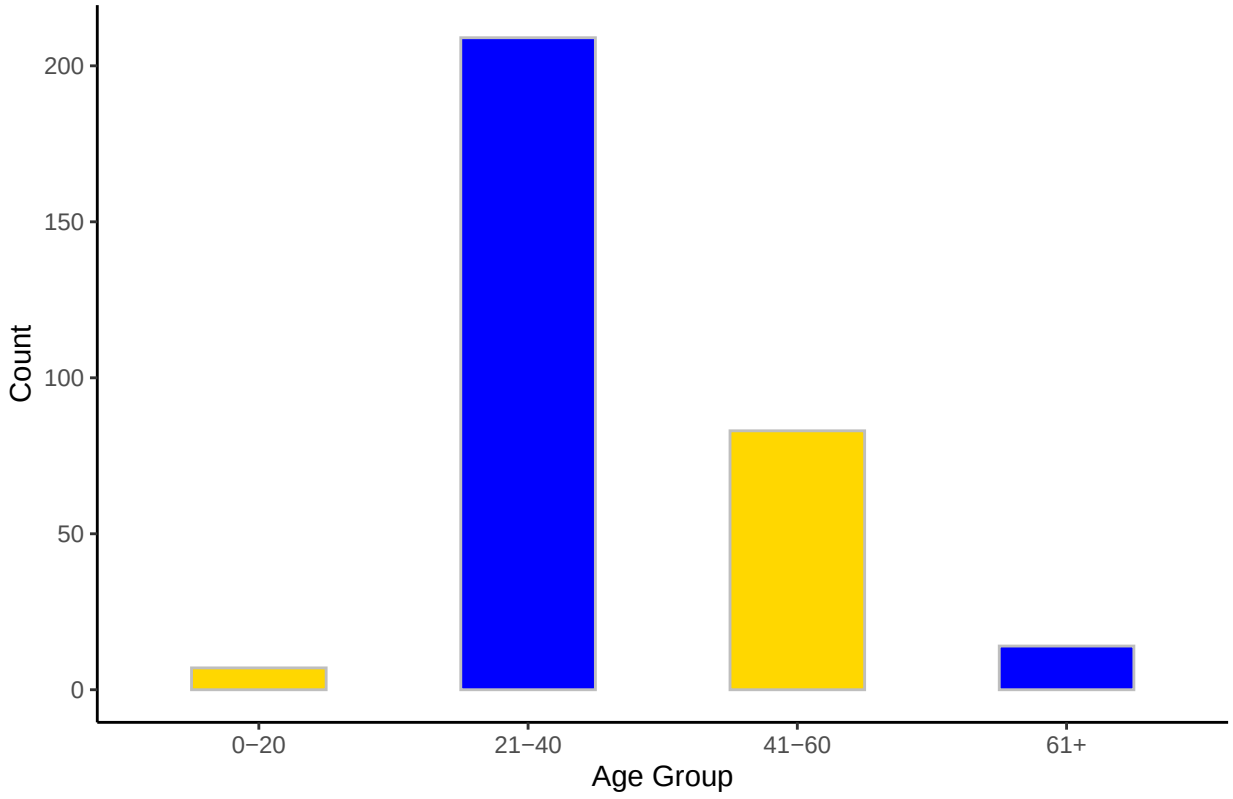


Figure 8. Education of survey respondents from experimenters' personal an

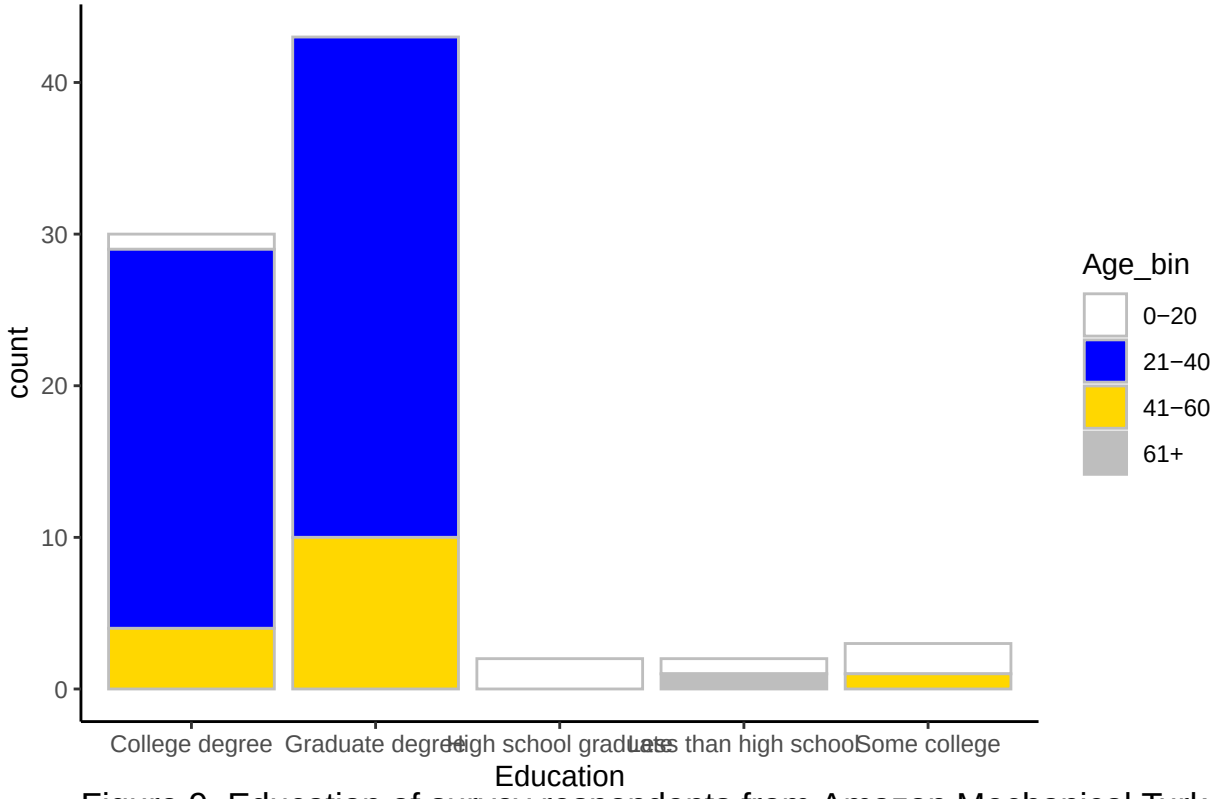



Figure 9. Education of survey respondents from Amazon Mechanical Turk

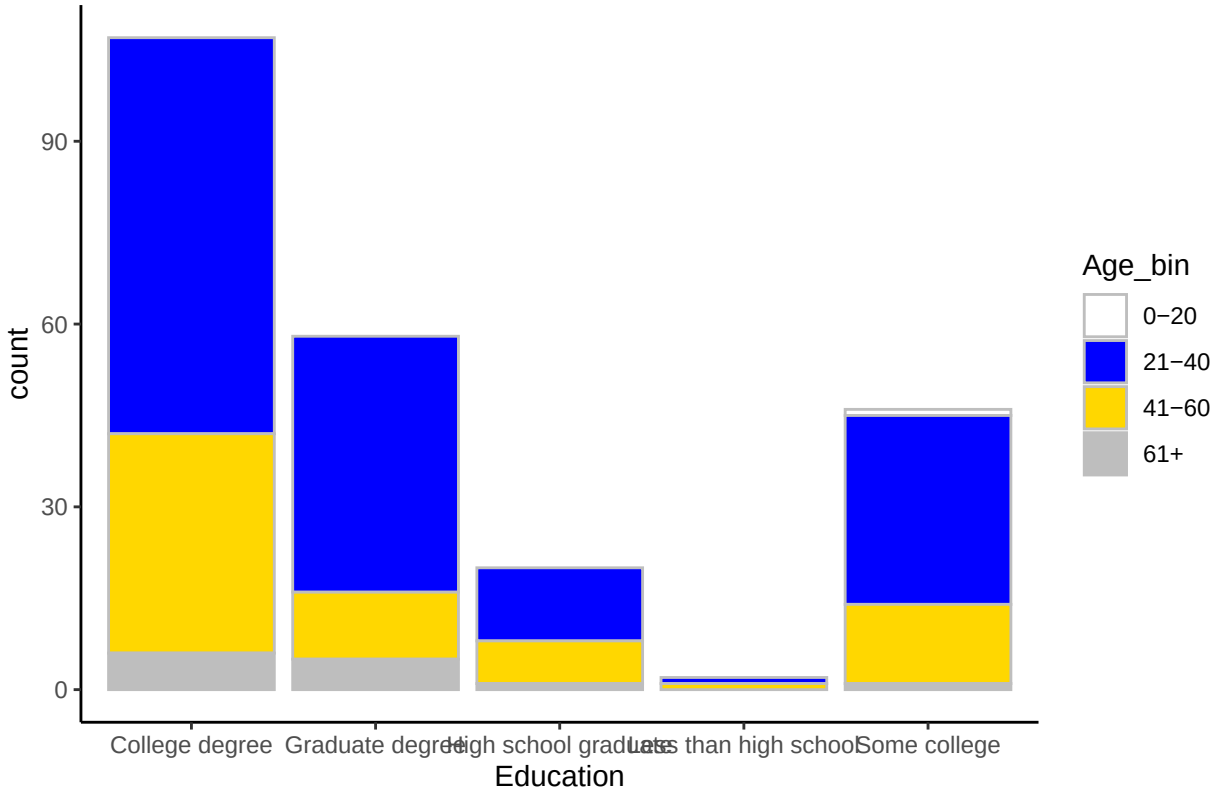


Figure 10. Mean scores by Age

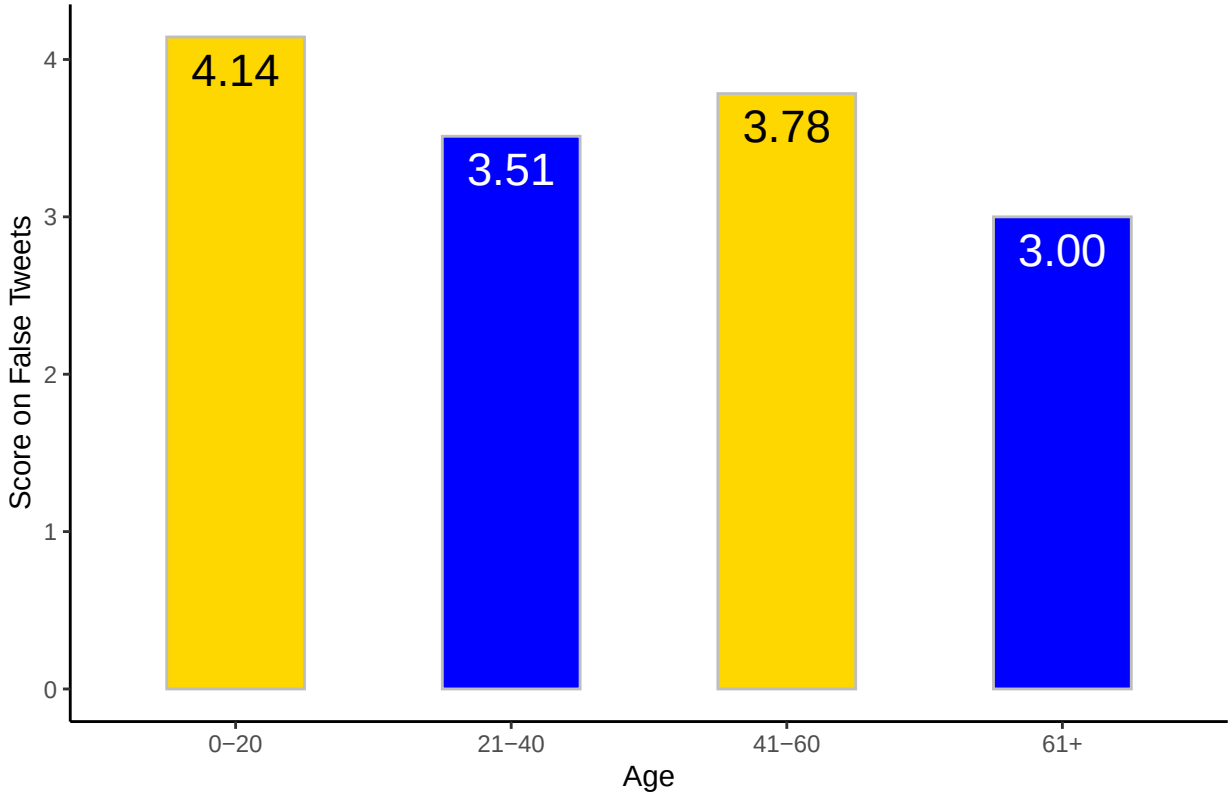
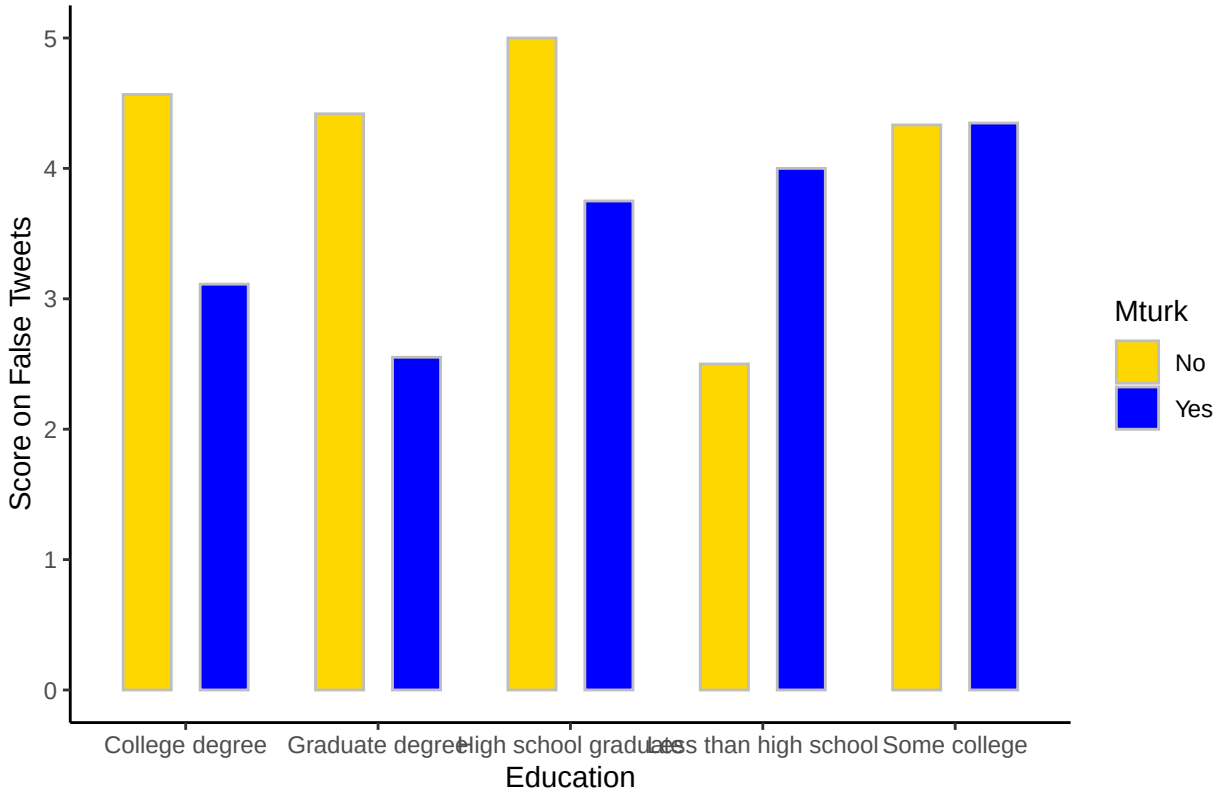



Figure 11. Mean scores by Education




```
## Saving 6.5 x 4.5 in image
```

```
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
## Saving 6.5 x 4.5 in image
```

Table 3, tests the interaction effects of treatment flag with these three co-variates to understand if warning about fake news effects demographics differently. Adding up the coefficients of interaction terms and applying Bonferroni correction, we find that neither education or age have any statistically significant marginal effects of warning flag on ability to detect false tweets.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:12

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:12

Now we proceed to test fixed question effects. Our survey has twitter posts *(tweets)* covering politics, science and general US current affairs. It is possible that there are marginal treatment effects based on the *article slant*. Our primary interest in determining the effect on score for identifying the false tweets correctly. Question numbers 3,5,6,9,10 contains false information in the survey. In this section, we regress the score of correct detection over the treatment and question. This effectively increases the amount of data in our data set since we have five questions for false and true posts respectively. Figure 12 illustrates this where we have aggregated the correct answer count for each question. The regression output is shown in Table 5. It shows a highly significant effect of the warning flag on participant's ability to detect false tweets [**0.18 (0.055)**]. This result suggests that there is an 18% improvement in score of false tweet detection when the question fixed effects are included for Question 3,5 and 6. It is worthwhile to remind ourselves that question 3 is regarding political stance regarding kneeling of players/coaches/staff when the national anthem is players during NFL games, to protest police brutality and racial profiling against African Americans in the United States. This topic has got considerable media and public attention and it is easy to understand how adding fixed effect for this topic's question generates a significant effect in our survey regression results. Questions 3 and 6 are not dealing with recent controversies but instead deal with a lasting and pervasive controvesy regarding scientific conspiracy and disbelief among the general US population. The results show that presence of a warning flag might be nudging participants to think twice about the tweets and the false contect that might be present in them.

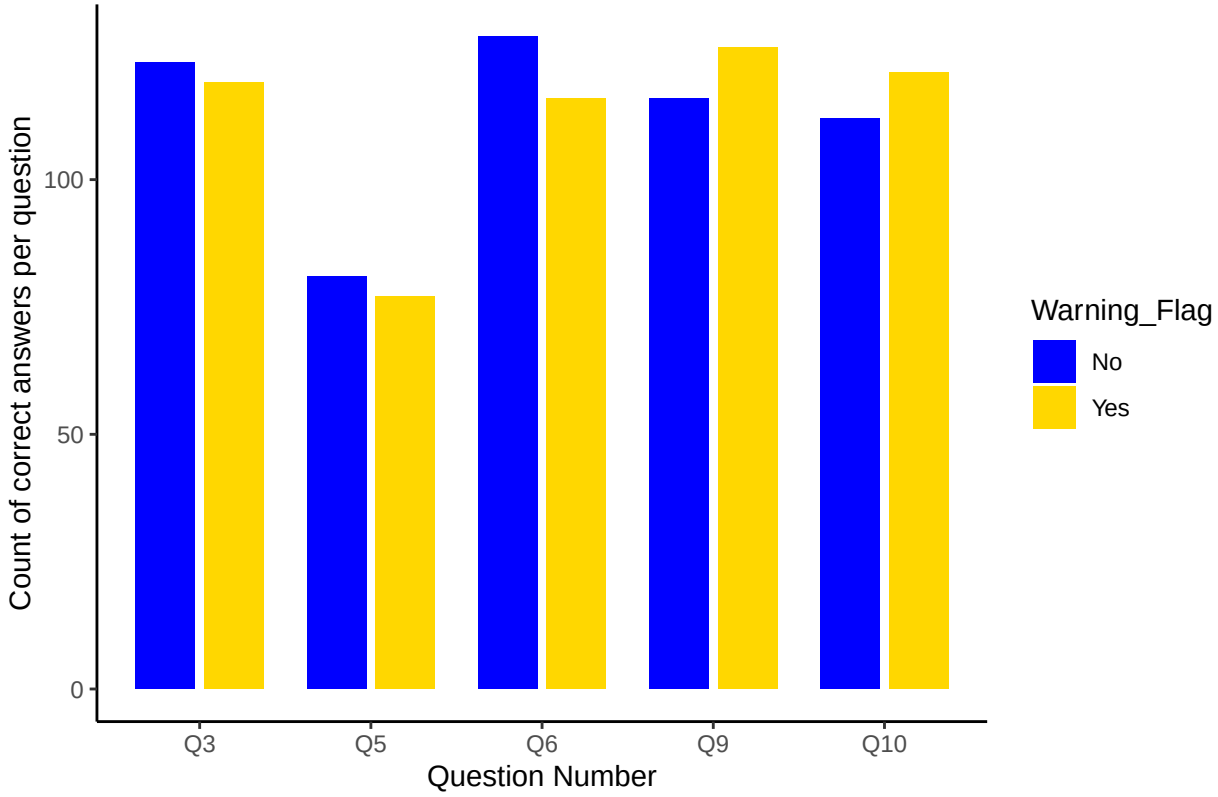
Table 3: Table 3. Treatment effect of warning flag on false tweet score with Age terms

| | Outcome Variable |
|---|---|
| | Score on False Tweets<br>Marginal effects of Age |
| Warning Flag | $-0.860^{***}$<br>(0.130) |
| MTurk Participant | $-0.440^{**}$<br>(0.180) |
| Captcha Verified | $-0.550^{**}$<br>(0.230) |
| Age_bin61+ | $-1.000^{*}$<br>(0.570) |
| Mturk | $-0.130$<br>(0.140) |
| captcha | $2.300^{***}$<br>(0.190) |
| assignment:Age_bin21-40 | $0.950^{***}$<br>(0.220) |
| assignment:Age_bin41-60 | $1.000^{***}$<br>(0.280) |
| assignment:Age_bin61+ | 0.740<br>(0.710) |
| Constant | $2.600^{***}$<br>(0.200) |
| Question Fixed Effects | No |
| Observations | 313 |
| $R^2$ | 0.460 |
| Adjusted $R^2$ | 0.440 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01<br>Regression models with robust standard errors. |

Table 4: Table 4. Treatment effect of warning flag on false tweet score with Education interaction terms

| | Outcome Variable |
|---|---|
| | Score on False Tweets<br>Marginal effects of Age |
| Warning Flag | −0.092<br>(0.210) |
| MTurk Participant | −0.390<br>(0.250) |
| Captcha Verified | 0.110<br>(0.280) |
| EducationLess than high school | −0.110<br>(0.730) |
| EducationSome college | 0.630***<br>(0.230) |
| Mturk | −0.450***<br>(0.140) |
| captcha | 2.100***<br>(0.190) |
| assignment:EducationGraduate degree | 0.550*<br>(0.320) |
| assignment:EducationHigh school graduate | 0.065<br>(0.450) |
| assignment:EducationLess than high school | −1.900<br>(1.300) |
| assignment:EducationSome college | 0.100<br>(0.340) |
| Constant | 2.500***<br>(0.260) |
| Question Fixed Effects | No |
| Observations | 313 |
| $R^2$ | 0.480 |
| Adjusted $R^2$ | 0.470 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01<br>Regression models with robust standard errors. |

Figure 12. General warning effect on scores for each false tweet




```
## Saving 6.5 x 4.5 in image
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:13

## Spillover Effects

Now we test formally the spillover effect of a warning flag on the participant's ability to detect true tweets correctly. Our alternate hypothesis was that there will be a negative effect on true tweet score in presence of a flag, i.e. when a general warning flag is present, subject will incorerctly label true tweets more often than when no warning about fake news is present. Table 6 shows that there is no statisitically significant [-0.180 (0.110)] negative effect of a warning flag on true tweet detection score. Table 7 shows statistically insignificant marginal effects of warning flag on all age groups: ATE on 21-40 [-0.2 (0.58)], 40-61 [-0.09 (0.74)], 0-21 [-0.35 (0.34) and 61+ [-0.65 (1.04)]. Simlilarly, Table 8 shows statisically insignificant marginal effects of warning flag on people in different education categories: Graduate Degree [0.13 (0.43)], College Degree [=-0.32 (0.17)], some college [0.1 (0.48)], high school graduate [0.03 (0.64)] and less than high school [-0.94 (1.27)]. Note, here we have once again scaled the standard errors to account for the number of interaction terms that are being considered in the regression to determine a statistical effect.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:14

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:14

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:14

Lastly, we test the spillover treatment effect with question fixed effects in Tables 9 and 10. We find from Table 10 that the only question that has a significant effect on the score of true tweet is Question 4 **[-0.23**

Table 5: Table 5. Treatment effect of warning flag on false tweet score with question fixed effects

| | Outcome Variable |
|---|---|
| | Score on False Tweets |
| Warning Flag | $0.180^{***}$ |
| | (0.055) |
| Q3 Correct | $-1.600^{***}$ |
| | (0.130) |
| Q5 Correct | $-1.300^{***}$ |
| | (0.063) |
| Q6 Correct | $-1.400^{***}$ |
| | (0.120) |
| Mturk Participant | $-0.120^{**}$ |
| | (0.054) |
| Captcha Verified | $0.230^{**}$ |
| | (0.096) |
| Constant | $4.700^{***}$ |
| | (0.110) |
| Question Fixed Effects | Yes |
| Observations | 313 |
| $R^2$ | 0.910 |
| Adjusted $R^2$ | 0.900 |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Regression models with robust standard errors.
Model contains question fixed effects for false tweets.
Effect is highly significant for a warning flag's ability to detect false tweets

Table 6: Table 6. Treatment effect of warning flag on true tweet score

| | Outcome Variable | | |
|---|---|---|---|
| | Score on True Tweets | | |
| | ATE(no indicator) | ATE(Mturk indicator) | Spillover Baseline(Mturk + Captcha |
| | (1) | (2) | (3) |
| Warning Flag | −0.180 | −0.180 | −0.180 |
| | (0.120) | (0.120) | (0.110) |
| MTurk Participant | | 0.340** | 0.042 |
| | | (0.140) | (0.150) |
| Captcha Verified | | | −0.690*** |
| | | | (0.130) |
| Constant | 3.600*** | 3.400*** | 4.100*** |
| | (0.084) | (0.130) | (0.180) |
| Question Fixed Effects | No | No | No |
| Observations | 313 | 313 | 313 |
| $R^2$ | 0.007 | 0.028 | 0.110 |
| Adjusted $R^2$ | 0.004 | 0.021 | 0.098 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Regression models with robust standard errors.
Warning Flag is the treatment and outcome variable is participant on true tweet questions in sur
Respondents in treatment saw the flag and those in control didn't see any warning.
Baseline mode is the third column in table with indicator variables for Mturk and BOT verificati

Table 7: Table 7. Treatment effect of warning flag on true tweet score with Age terms

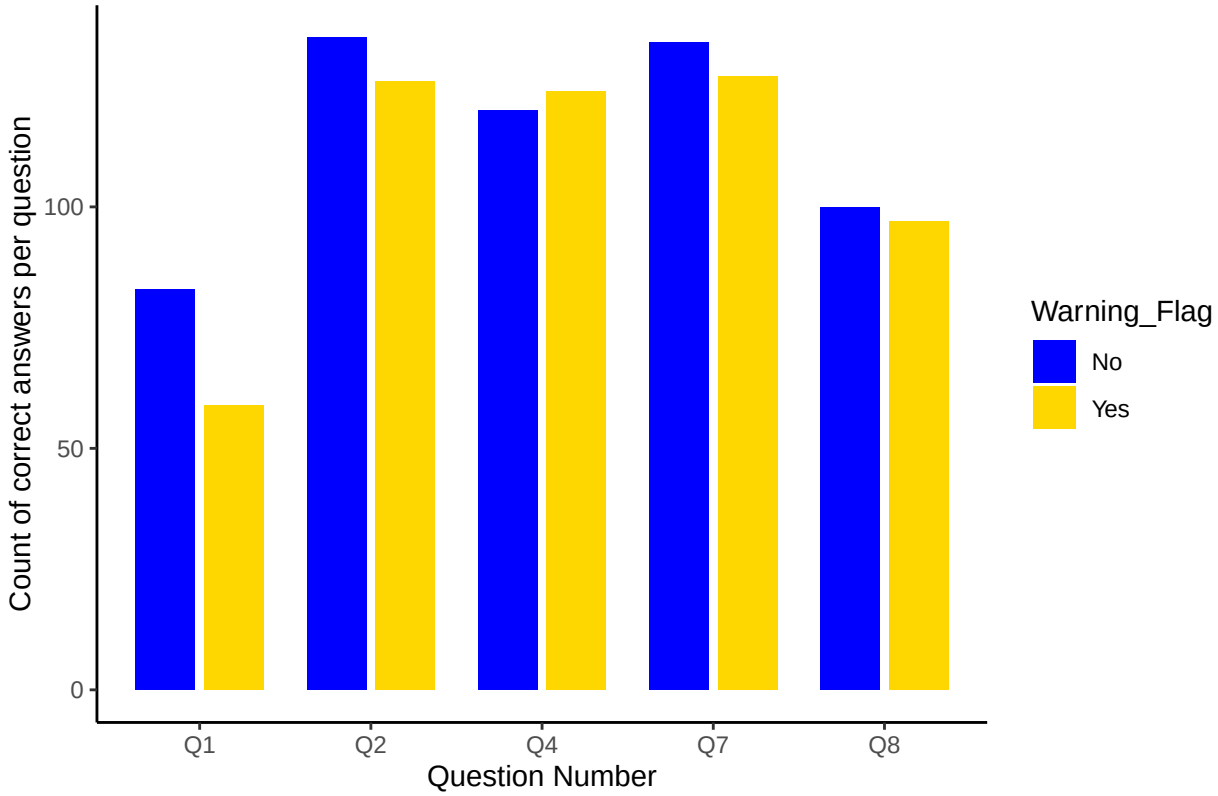
| | Outcome Variable |
|---|---|
| | Score on True Tweets<br>Marginal effects of Age |
| Warning Flag | 0.350<br>(0.340) |
| Age (21-40) | 0.071<br>(0.240) |
| Age (41-60) | −0.100<br>(0.280) |
| Age (61+) | 0.130<br>(0.440) |
| Mturk Participant | 0.092<br>(0.160) |
| Captcha Verified | −0.660***<br>(0.140) |
| Warning Flag X Age (21-40) | −0.550<br>(0.370) |
| Warning Flag X Age (41-60) | −0.440<br>(0.400) |
| Warning Flag X Age (61+) | −1.000<br>(0.700) |
| Constant | 4.000***<br>(0.240) |
| Question Fixed Effects | No |
| Observations | 313 |
| $R^2$ | 0.120 |
| Adjusted $R^2$ | 0.089 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01<br>Regression models with robust standard errors. |

Table 8: Table 8. Treatment effect of warning flag on true tweet score with Education interaction terms

| | Outcome Variable |
|---|---|
| | Score on True Tweets<br>Marginal effects of Age |
| Warning Flag | $-0.320^{*}$<br>(0.170) |
| Graduate Degree | $-0.140$<br>(0.190) |
| High School Graduate | $-0.430$<br>(0.350) |
| Less than High School | $-1.100^{***}$<br>(0.380) |
| Some College | $-0.360^{*}$<br>(0.210) |
| Mturk Participant | 0.057<br>(0.150) |
| Captcha Verified | $-0.630^{***}$<br>(0.130) |
| Warning Flag X Gradute | 0.190<br>(0.260) |
| Warning Flag X High School | 0.350<br>(0.470) |
| Warning Flag X Less than high school | $-0.620$<br>(1.100) |
| Warning Flag X Some College | 0.420<br>(0.310) |
| Constant | $4.100^{***}$<br>(0.230) |
| Question Fixed Effects | No |
| Observations | 313 |
| $R^2$ | 0.140 |
| Adjusted $R^2$ | 0.110 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01<br>Regression models with robust standard errors. |

**(0.099)**]. Question 4 in the survey asked participants about a tweet containing data, from National Center for Education Statistics, about proficiency of 8th graders in US history. The tweet said that only 15% of 8th graders are proficient in US history (which is a true fact). The presence of a warning flag seems to reduce people's ability to correctly identify this statistic as true by an average 23% (SE: 9.9%) compared to when no warning flag about fake news was presented to survey takers.

Figure 12. General warning effect on scores for each true tweet




```
## Saving 6.5 x 4.5 in image
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Aug 09, 2020 - 15:34:15

Table 9: Table 3. Treatment effect of warning flag on true tweet score with question fixed effects

| | Outcome Variable |
|---|---|
| | Score on True Tweets |
| Warning Flag | $-0.230^{**}$ |
| | (0.099) |
| Q4 Correct | $1.100^{***}$ |
| | (0.130) |
| Mturk Participant | $-0.009$ |
| | (0.130) |
| Captcha Verified | $-0.710^{***}$ |
| | (0.110) |
| Constant | $3.300^{***}$ |
| | (0.190) |
| Question 4 Fixed Effects | Yes |
| Observations | 313 |
| $R^2$ | 0.300 |
| Adjusted $R^2$ | 0.290 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |
| | Regression models with robust standard errors. |
| | Model contains question fixed effects for true tweets. |
| | Effect is significant with question 4 fixed effect |