Does credibility in the scientific community translate to credibility in the general public?

Jackson Argo, Oscar Casas, Gabriel Louis-Kayen, Peter Morgan

04/07/2022

Abstract

Does credibility in the scientific community translate to credibility in the general public? There's growing phenomena of skepticism and outright rejection of scientific findings across the world, critically dubbed the Anti-Science movement. The common hypothesis is that people are generally less likely to accept peer-reviewed, published scientific findings and more likely to accept anecdotal findings of like-minded individuals. Some of this phenomena is attributed to a Fake News crisis, where it has become increasingly easy to spread misinformation over social media platforms. The assumption in research regarding Fake News is that the creators are intentionally misleading people to believe that the fake news is in fact credible. However, we posit that distrust of the scientific community has become so pervasive that it is instead the association with a peer-reviewed scientific journal that makes new findings less credible to the general public.

Contrary to our hypothesis, we discovered that the perception of peer review increased the trustworthiness of claims. After running our experiment, we found that those who received treatment were around 10% more likely to trust the claim. This result was statistically significant at a 95% confidence level.

Experimental Design

Experimental Overview

We created summaries of two research articles about UV radiation, and survey respondents were asked to read the two summaries and decide which one seemed more trustworthy. We designed the two summaries to provide very similar and reasonable information, but without any specific information about the source of the data, aside from saying that these are findings from news articles. The treatment and control groups both received the same first article, but for the second article, we added "According to a peer reviewed study," to the beginning of the first sentence in order to inform the reader what they read did in fact come from a peer reviewed, scientific article. In order to check that the respondents actually read the article, we also included a multiple choice question that asks them what the articles are about. Finally, we included an open ended question that asks them for a brief reason why they made their choice.

Table 1: Articles given to subjects in the survey.

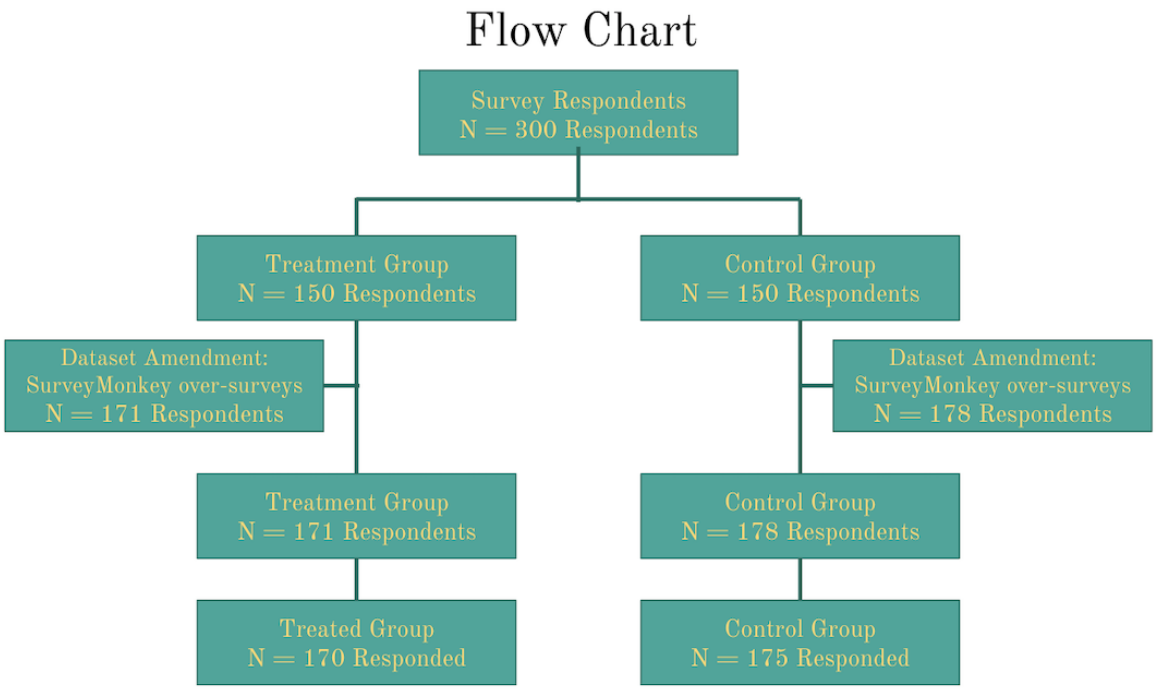
Article 1: It is well known that chronic exposure to ultraviolet (UV) radiation present in sunlight is responsible for the induction of most nonmelanoma skin cancer (NMSC) in humans. Wavelengths in the UV-B (290-320 nm) region of the solar spectrum are absorbed into the skin, paroducing erythema, burns, and eventually skin cancer.
Article 2 (Control): Too much UV radiation from the sun or sunbeds can damage the DNA in our skin cells. DNA tells our cells how to function. If enough DNA damage builds up over time, it can cause cells to start growing out of control, which can lead to skin cancer. Anyone can develop skin cancer, but some people can have a higher risk, including people who burn more easily.
Article 2 (Treatment): According to a peer reviewed study, too much UV radiation from the sun or sunbeds can damage the DNA in our skin cells. DNA tells our cells how to function. If enough DNA damage builds up over time, it can cause cells to start growing out of control, which can lead to skin cancer. Anyone can develop skin cancer, but some people can have a higher risk, including people who burn more easily.

Randomization

Randomization occurs at the individual level by Survey Monkey. We have two survey options for subjects to take, one for control and one for treatment. As the subject clicks to begin the survey, they are randomly assigned to either treatment or control groups. This means there is no guarantee that the treatment and control groups are the same size because the chance of being in treatment and control is 50% for each subject independent of other subjects assignments.

In our survey, subjects always read article A first and when answering the question of which article they trust more, article A is always presented as the first option. In hindsight, these are two things we would have randomized because we are going to have some respondents select the first option through the whole survey. This could cause an increase in the number of subjects that will select article A, however, the estimate we obtain is not biased. This is because we are using a difference in differences method to calculate our outcome and since this problem exists in the treatment and control groups, the difference in differences makes the estimate unbiased. If we did randomize the order of the articles and questions, this would remove some noise in our findings.

We directed Survey Monkey to sample 300 individuals, 150 in the control group and 150 in the treatment group. But because Survey Monkey sends the survey to more potential respondents than requested (in case of non-respondents), our control survey received 178 respondents and our treatment survey received 171 respondents. After very low levels of attrition, our control group ended up having 175 individuals and our treatment group ended up having 170 individuals. The following flowchart shows our randomized survey assignment and response process.

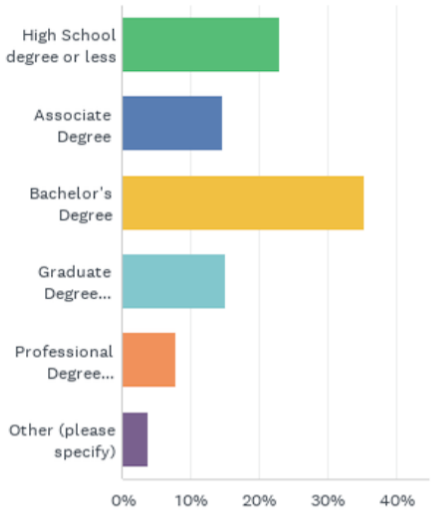


Distributions

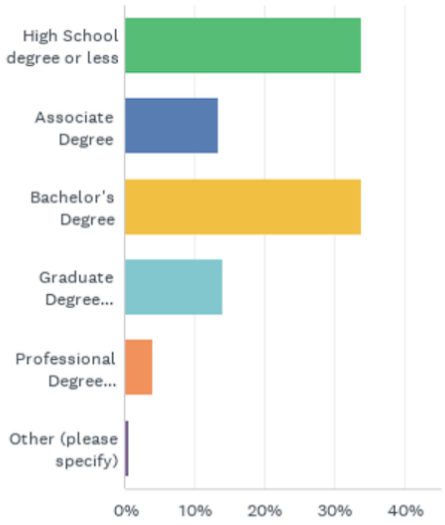
We requested that Survey monkey survey individuals that are of any gender, between the ages of 18 and 70, with any level of education and across all regions of the Unites States. Survey Monkey even provided information on the distribution of income of the control and treatment groups. This led to the following distributions. As seen below, the control group and treatment group have very similar distributions for income level and for education. And while the distributions for age take different shapes, with the control distribution having a left-skew and the treatment distribution having a right-skew, all of the age bins have a considerable number of respondents in both groups.

Education

Control

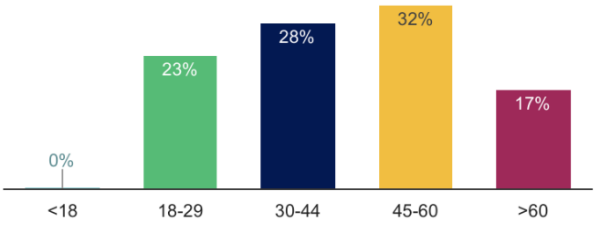


Treatment

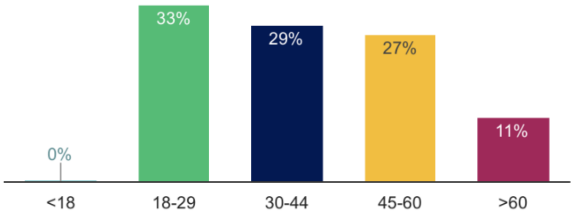


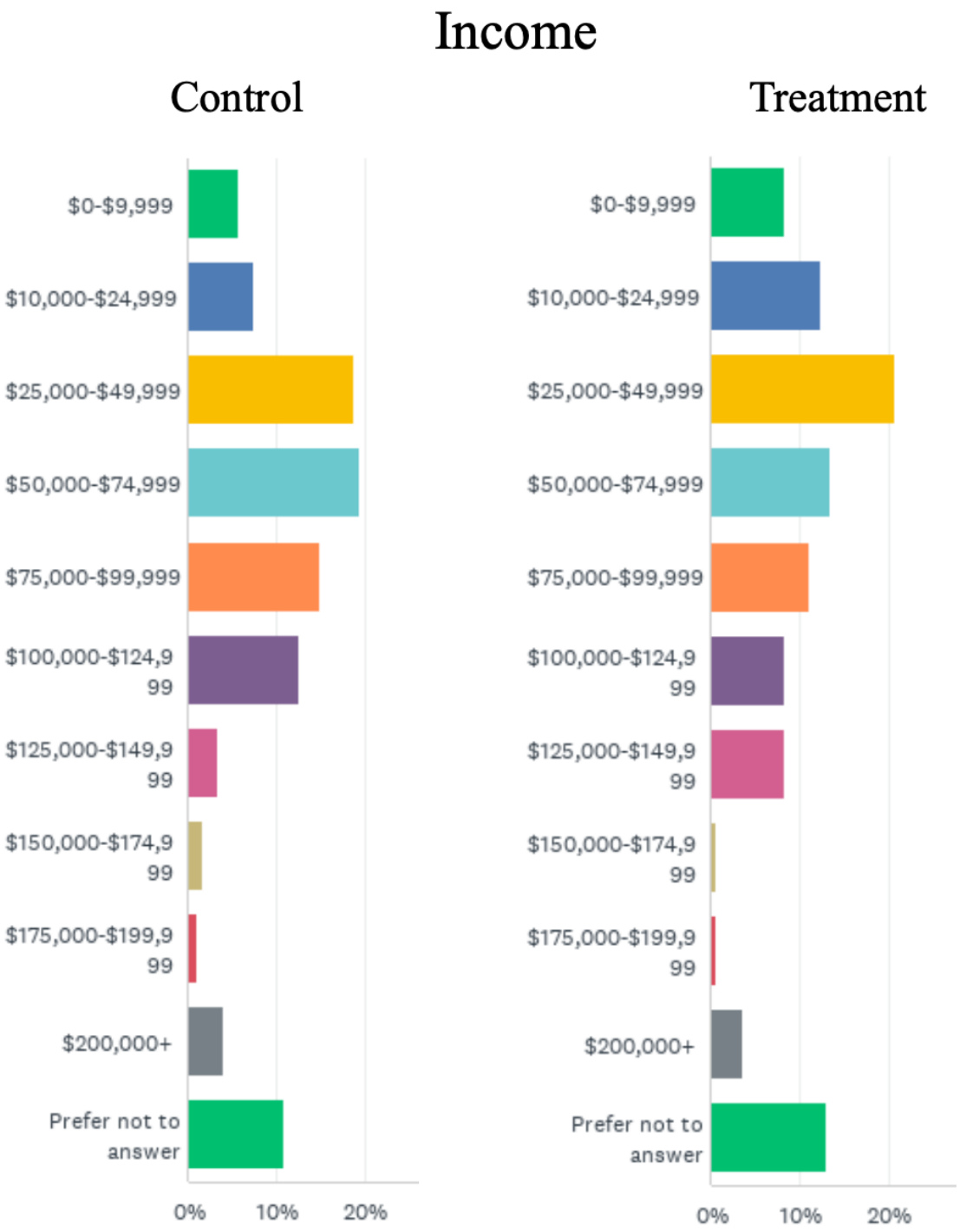
Age

Control



Treatment

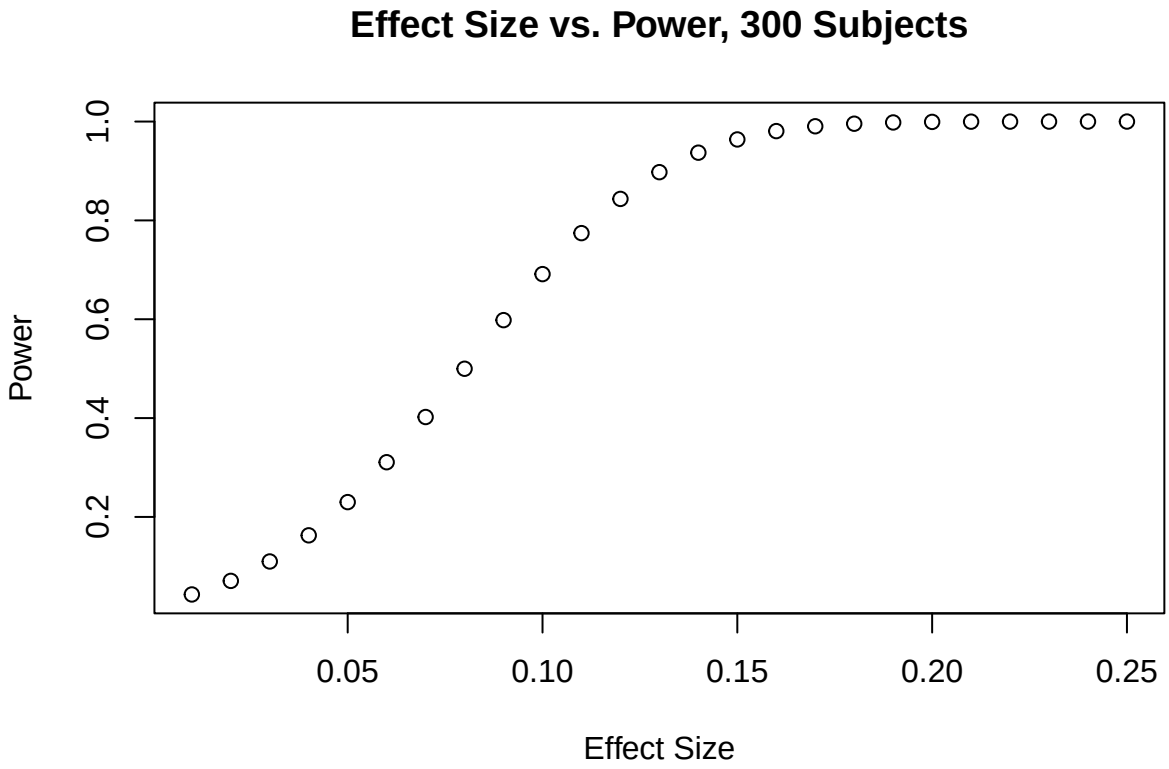




Power Analysis

For our power analysis we first considered the number of subjects we want to have. We will consider 300 subjects because we feel we can get decent power at this number and it also leaves us with room in the budget if more subjects are needed. To calculate power, we first start by evenly splitting our 300 hypothetical subjects evenly into control and treatment. For our control group, we assume that half of our subjects will report that article A is more trustworthy, and half of our subjects will report that article B is more

trustworthy. We then iterate through different effect sizes starting from 0.01 up to 0.25 by increments of 0.01. Power is calculated for each effect size and the results are plotted below.



As you can see from the plot above, we get reasonable power when our effect size is larger than 0.10. Depending on our results, we may need to recruit more subjects. An example of this would be if at 300 subjects there is an effect of 0.075 and a p-value of 0.07. This result is close to significant but still not significant. Our power calculations show that at an effect size of 0.075, our power is only around 0.5. This means that even if there is a true difference between treatment and control, there is only a 50% chance that we would properly reject the null hypothesis. Our strategy would then be to purchase more subjects as there is a higher chance that we would properly reject the null hypothesis.

Results

The initial results show a large shift in the number of people who found Article B most trustworthy between treatment, 52.3529412, and control 41.7142857. With such a dramatic shift, we did find statistically significant results, which we will go into in more detail in the Model Results section.

Why respondents made their choice.

In addition to choosing which article is most trustworthy, we also asked respondents to give a brief reason why they made their choice. The answers were free response, and most respondents gave a few words to a sentence long explanation. We categorized the responses into 7 groups: better explanation, both trustworthy, both unstrustworthy, citation, first sentence, more reasonable, and more technical. The groups both trustworthy and both untrustworthy indicate that respondents weren't really sure which one to chose. Relatively few respondents fell into these two categories, 3 found both trustworthy, and 7 found both trustworthy. The more reasonable category is used for respondents that indicated the article they chose was more believable or seemed more logical.

The first sentence category represents respondents who found the first sentence of Article A, starting with "It is well known that," was off-putting and that, perhaps, it is not well known that UV radiation causes skin cancer. This in turn caused these respondents to find Article B more trustworthy. The citation category represents respondents who noted that their choice contained a reference or citation regarding scientific literature. Only the treatment version of Article B contained such a reference, however several respondents in control said they chose Article A for this reason.

The categories better explanation and more technical represent opposite sentiments about what makes an article more trustworthy. Respondents typically found Article B to be a better explanation, meaning it was easier to understand or more accessible to a general audience. In contrast, respondents typically found Article A to be more technical; it used more scientific terminology, but was less easy to comprehend. Within these two categories, there was typically a mention that the other article was less trustworthy because it felt into the opposing category. For example, respondents would say that Article B provides a better explanation, and thus is more trustworthy than the overly technical Article A. Similarly, respondents would say that Article A feels like it was written by a scientist, as opposed to Article B, which felt dumbed down.

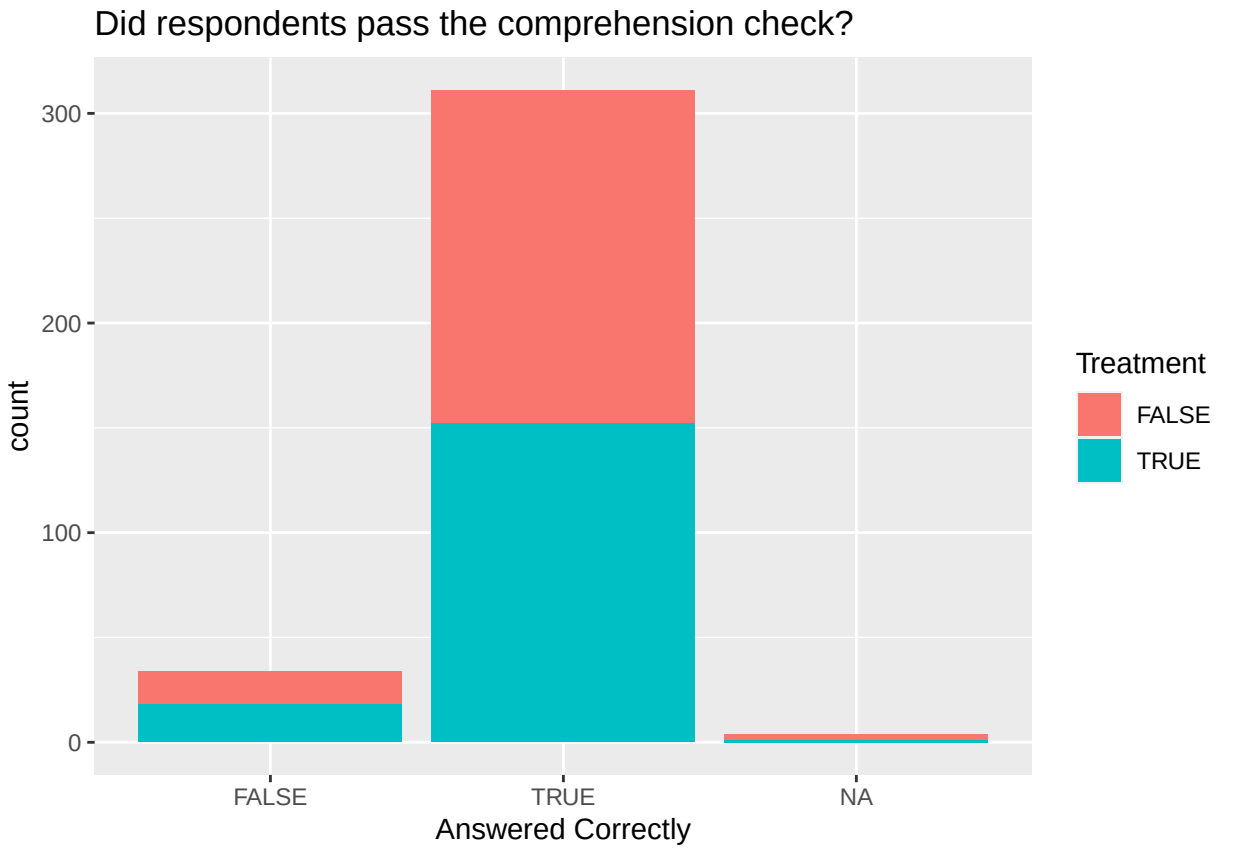
Attrition Analysis

In total, we had 4 respondents that answered none of the questions, 1 from treatment and 3 from control. We take this to mean that some percentage of respondents in general will not answer any of the questions on the survey. We performed a binomial test to verify that this assumption. The binomial model uses the ratio of control respondents who answered the research question as the success probability, 0.994152, and tests that this ratio is similar to the ratio of treatment respondents, 0.9831461. This test returns a 95% confidence interval of [0.951538, 0.9965107] ($p = 0.0874436$). The confidence interval includes the ratio for control respondents, and gives evidence that there is no differential attrition.

Non-compliance

In this experiment, we attempted to use a comprehension check to determine if subjects were compliers. We considered a complier as someone who read the articles that were given to them in the survey. We asked all subjects what the articles were about as a comprehension check to determine if subjects actually read the articles. Subjects were given 4 multiple choice options. The correct option was "The effect of UV radiation on developing skin cancer". The three other options were "Micro-plastics in the Pacific Ocean", "Rates of deforestation in the Amazon Forest", and "How vaccines effect the rate of COVID-19 infection".

In the treatment group, 152 out of 170 subjects (89.41%) answered this question correctly. In the control group, 159 out of 175 subjects (90.86%) answered this question correctly. These results are shown in the plot below with the attriters being shown in the NA column.



Initially, our plan was to only use the subjects who passed the comprehension check and calculate a complier average causal effect. We decided against this because we believe there is a problem that could cause this measure of non-compliance to not be accurate. Since there are 4 multiple choice options, there is a chance that people who did not read the articles get the comprehension check correct. Our initial thought to account for this was to divide the number of incorrect answers by 0.75 to get the the true non-compliance rate. The problem with this is it makes the assumption that non-compliers' answers to the comprehension check are uniformly dsitributed which is not true. We found that for treatment and control, only 1% of subjects chose the micro-plastic option while the other incorrect options were picked between a 3% and 5% rate. This means we cannot make any assumptions on the rate of non-compliers that choose the correct answer to the comprehension check. Because of this, we have decided to report our results as the intent to treat effect instead of trying to calculate a complier average causal effect.

How similar are the control articles?

In our design, we expected to see an almost equal distribution between control group's choice of articles. However, there was a strong preference for the first article, Article A. Again, we use a binomial test check for a 50/50 preference. We subset the data to only include control members who submitted a response. The binomial test returns a 95% confidence interval of [0.506059, 0.6567996], which barely does not include 0.50. The articles were not shown in a random order, and Article A was always shown first, followed by Article B. There could be respondents who always chose the first answer, which could explain the strong preference for Article A.

Model results

For the first model, we ran a regression of only group assignment on whether the the respondents chose Article B as the most trustworthy article, using robust standard errors. We found that the treatment did have a statistically significant effect of 0.106387 (0.053760), meaning that the short phrase "According to a

peer reviewed study" did the make the article seem more trustworthy. This is against our original hypothesis, however it is a relief in terms of scientific credibility.

In the second model, we included the comprehension check as a covariate. The comprehension check is coded as a binary variable to indicate whether the respondent answered correctly to the question "What were the articles about?". Most respondents did answer this correctly, and including this term had a marginal affect on the coefficient and error term for group assignment 0.107376 (0.053883).

Finally, we produced a third model that contains all the group assignment, comprehension check, and the covariate data provided by Survey Monkey. This model showed a small increase in the coefficient and improvement in the error term for group assignment, 0.1145013 (0.048367).

We also produced a fourth model, containing the same covariates as before, in addition to covariates to represent the reason why people made their choice. The error term did increase in this fourth model, and we believe this is because including the respondent's reason is actually a bad control. Specifically, people who said that the citation is why is almost a direct result of being assigned to the treatment group. The reason they made their choice then is directly dependent on the outcome variable, which means that we've included a dependent variable in our regression. This model is not included in the model results table, but is included in the appendix, for completeness sake.

Table 2: Model Results

	Dependent variable:		
	Choose Article B		
	(1)	(2)	(3)
Treatment	0.106** (0.054)	0.107** (0.054)	0.115** (0.058)
Understood article		0.068 (0.092)	0.046 (0.102)
Age			0.005** (0.002)
Male			0.042 (0.062)
High School degree			−0.015 (0.223)
Associate degree			−0.124 (0.232)
Bachelor's degree			−0.075 (0.224)
Graduate degree			−0.039 (0.232)
Professional degree			0.098 (0.254)
Device: iOS			0.059 (0.061)
Device: MacOS			−0.029 (0.533)
Device: Windows			0.118 (0.176)
Region: East North Central			0.088 (0.300)
Region: East South Central			−0.059 (0.315)
Region: Middle Atlantic			0.221 (0.301)
Region: Mountain			0.170 (0.315)
Region: New England			−0.036 (0.323)
Region: Pacific			0.173 (0.299)
Region: South Atlantic			0.017 (0.297)
Region: West North Central			0.192 (0.313)
Region: West South Central			0.199 (0.303)
Income: 0-9,999			0.036 (0.139)
Income: 10,000-24,999			0.058 (0.126)
Income: 25,000-49,999			−0.025 (0.112)
Income: 50,000-74,999			−0.039 (0.117)
Income: 75,000-99,999			0.027 (0.124)
Income: 100,000-124,999			−0.033 (0.129)
Income: 125,000-149,999			0.112 (0.154)
Income: 150,000-174,999			0.479*** (0.170)
Income: 175,000-199,999			−0.231 (0.537)
Income: 200,000+			0.122 (0.168)
Constant	0.417*** (0.037)	0.355*** (0.091)	0.028 (0.374)
Observations	345	345	345
R^2	0.011	0.013	0.110
Adjusted R^2	0.008	0.007	0.022
Residual Std. Error	0.498 (df = 343)	0.498 (df = 342)	0.494 (df = 313)
F Statistic	3.940** (df = 1; 343)	2.258 (df = 2; 342)	1.245 (df = 31; 313)
Note:			*p<0.1; **p<0.05; ***p<0.01

Conclusions

Our experiment produced both practically significant and statistically significant results on how the use of peer-reviewed sources affects the credibility of information. The emergence and rapid expansion of the Anti-Science movement across the United States led us to hypothesize that findings that are associated with peer-reviewed scientific journals are typically found to be less credible among the general public. To test this hypothesis, we conducted a difference-in-differences research experiment by surveying nearly 350 individuals of all genders, between the ages of 18 and 70, and residing in every region of the United States. Analyzing our survey results, the data we collected led us to reject our null hypothesis that information associated to peer-reviewed scientific journals has no effect on credibility among American residents that fit within the demographics that we surveyed. Even as we included more covariates in our regression to explain some of the variation in the credibility difference between the two articles we included in our survey, all 4 of our regression models produced statistically significant findings. Each of these models indicated that association to a peer-reviewed journal increases the credibility of information by over 10% with an average increase of 10.85% across all 4 models.

While the experiment does not ultimately offer any conclusions on the effect of the Anti-Science movement, it does show the value of the peer-review scientific process. The peer-review process does not only support findings of the scientific community, but helps much of the American public digest and believe new information. Indicating when information comes from a peer-reviewed source could have massive effects across a large audience. Presenting trustworthy information could lead individuals to choose better medical decisions, to adopt healthier behaviors, to make more financially beneficial choices, among countless ways to improve indiviuals' livelihoods. Especially in regards to the explosion of anti-science sentiment in the wake of the COVID-19 pandemic, offering peer-reviewed information on public health could (and likely already does) save lives.

Limitations / Future Follow Up Studies

Our research experiment suffers from multiple limitations that may bias our results or make our findings less generalizable. First and foremost, the articles we used in our surveys only discuss the effects of UV radiation on skin damage. Information regarding this topic may be more (or less) widely regarded as credible than information regarding a different topic. The effect of association to a peer-reviewed source may be drastically different among scientific topics that are more commonly contentious and that are more relevant in the Anti-Science movement, such as the efficacy of vaccines and how human behavior affects the global climate. Topics like these may even see less perceived credibility in information that is associated to peer reviewed sources. Ultimately, this may mean that our findings are biased by the scientific topic we chose to present in our surveys and that our findings can not be generalized beyond scientific topics that are similar to the topic of UV radiation and skin damage. It would be interesting to perform the same study using articles from different topics to see how the results vary.

Additionally a binomial test we performed showed that the articles we presented had a statistically significant difference in credibility in the control group. This may mean that the difference in the content of each article (not including the treatment) may distort the true effect of sourcing information as peer-reviewed. As mentioned earlier, even the ordering of the Article A to always appear before Article B may have made our results more noisy. An immediate correction that we would take in performing this experiment again would be to ensure that the order of the articles in the survey was random.

While we did address and limit the effect noncompliance could have, the fact still remains that noncompliance could have played a role in the study. We defined noncompliance as participants that did not read the study or randomly answered questions, to mitigate this we had a comprehension check which the majority of people answered correctly. That being said due to it being a multiple choice question it leaves space for noncompliers to count as compliers. It also poses the question if they actually received treatment. Many people can skim an article and answer the comprehension check correctly, but we cannot be certain that they readily identified the signal for peer review.

Another key component we would like to gather information on is attrition. We had a low attrition rate and

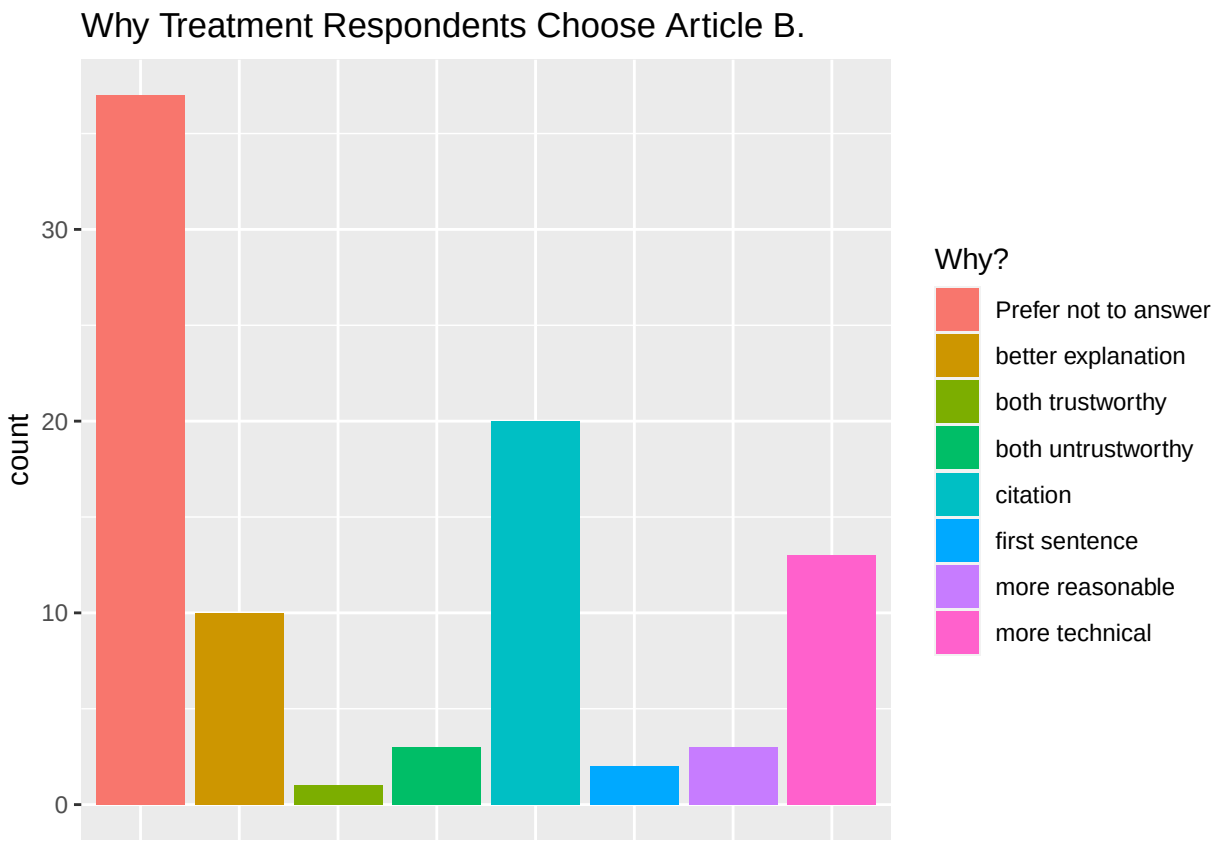
we expect the reason for the attrition to not be correlated with the effect of treatment. However, due to the fact that survey monkey does not give data on if the attrition was pre or post treatment, we would like to know what caused the attrition and if by any chance it has a correlation with the effect of treatment.

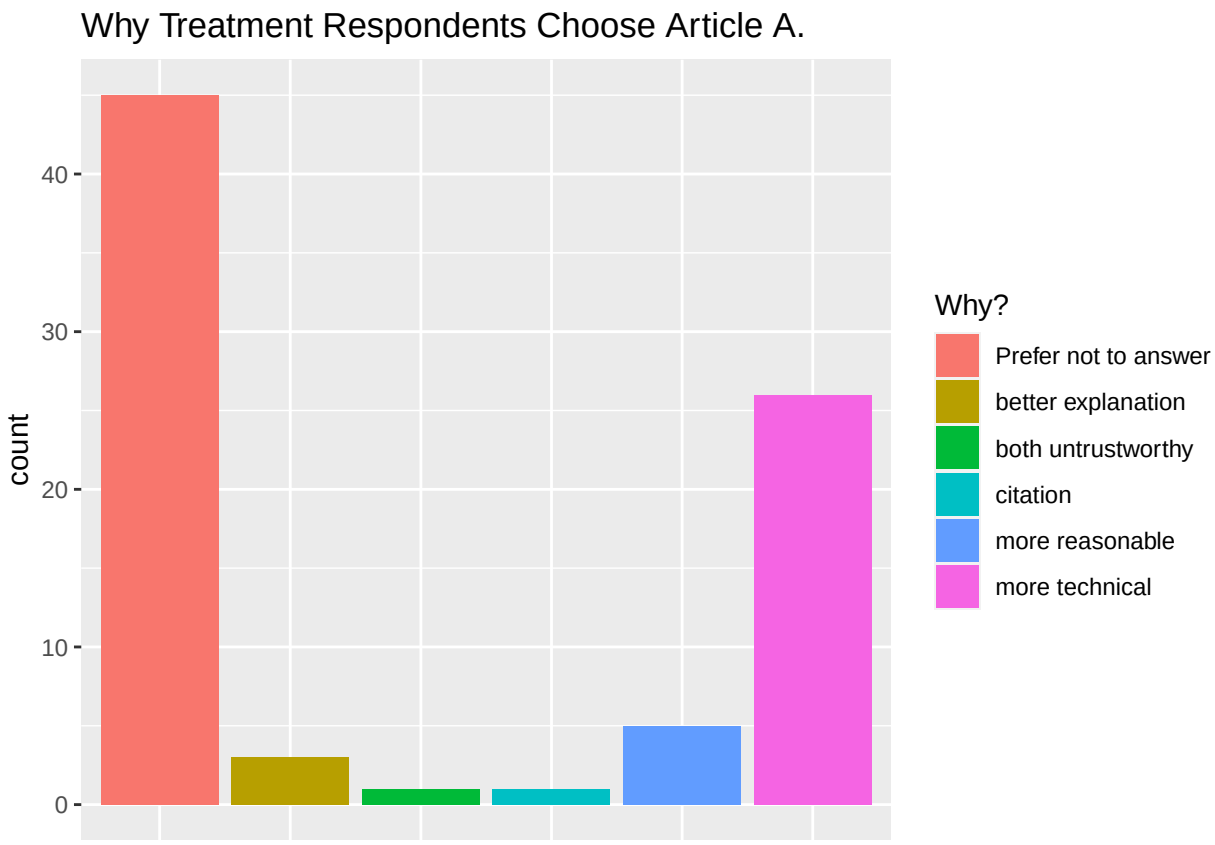
Lastly, our initial hypothesis and our decision to even conduct this research experiment were informed by the resent surge of anti-science sentiment both in the United States and around the world. While our experiment's results offer a better understanding of how peer-reviewed information is trusted, it does not actually offer any conclusive insight into the effects of the Anti-Science movement. It would be interesting to more narrowly perform this experiment in communities that regularly consume news that reinforces trust in the scientific community and in communities that regularly consume news that spreads distrust in the scientific community to observe the differences between these opposing communities (a difference in difference in difference experiment if you will).

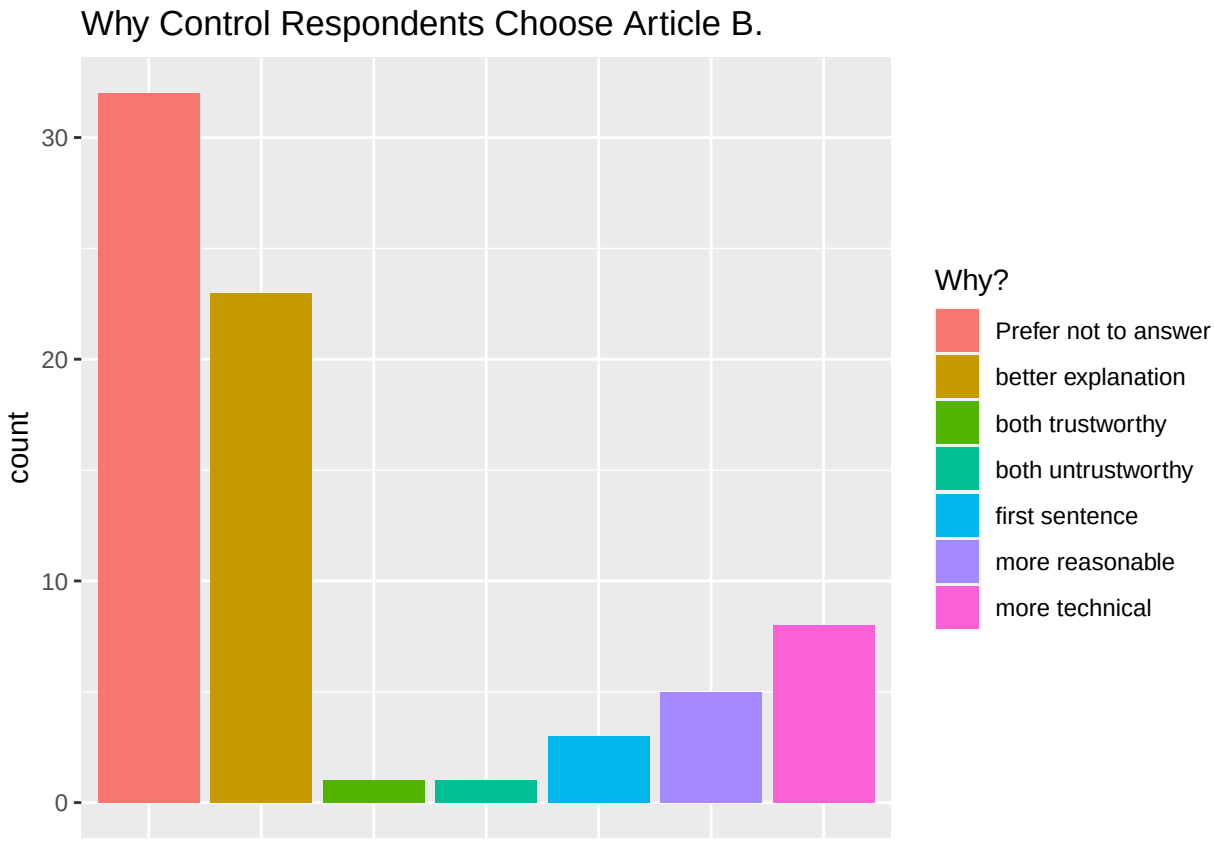
Appendix

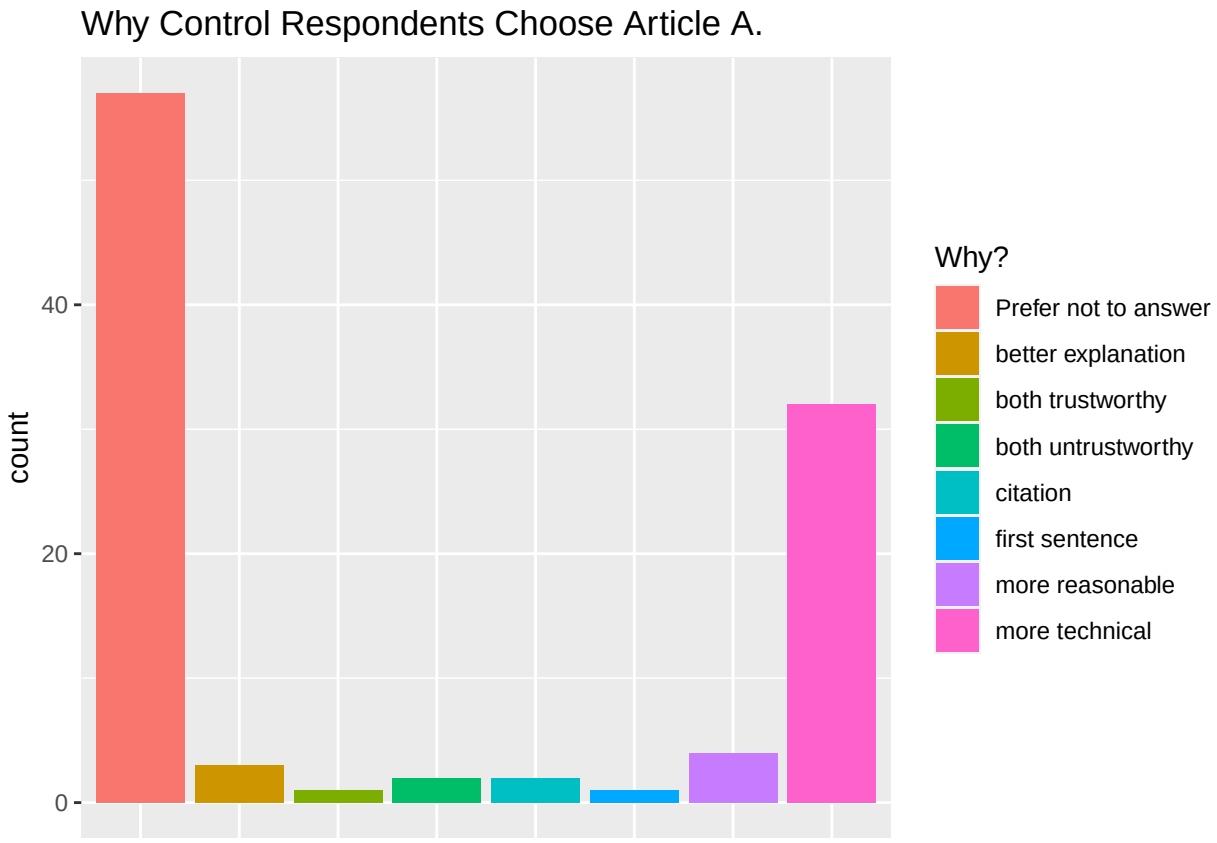
Table 3: Bad Controls Model

	Dependent variable:
	Choose Article B
Treatment	0.106* (0.058)
Understood article	−0.010 (0.105)
Age	0.005** (0.002)
Male	0.054 (0.058)
High School degree	−0.094 (0.193)
Associate degree	−0.235 (0.199)
Bachelor's degree	−0.145 (0.193)
Graduate degree	−0.158 (0.202)
Professional degree	−0.021 (0.230)
Device: iOS	0.065 (0.057)
Device: MacOS	0.118 (0.621)
Device: Windows	0.109 (0.171)
Region: East North Central	0.223 (0.216)
Region: East South Central	0.083 (0.239)
Region: Middle Atlantic	0.347 (0.220)
Region: Mountain	0.214 (0.241)
Region: New England	0.120 (0.241)
Region: Pacific	0.293 (0.218)
Region: South Atlantic	0.148 (0.214)
Region: West North Central	0.289 (0.237)
Region: West South Central	0.282 (0.221)
Income: 0-9,999	0.077 (0.133)
Income: 10,000-24,999	0.052 (0.118)
Income: 25,000-49,999	−0.045 (0.101)
Income: 50,000-74,999	−0.027 (0.106)
Income: 75,000-99,999	0.005 (0.110)
Income: 100,000-124,999	−0.070 (0.114)
Income: 125,000-149,999	−0.070 (0.162)
Income: 150,000-174,999	0.481* (0.252)
Income: 175,000-199,999	−0.290 (0.651)
Income: 200,000+	0.094 (0.148)
Reason: Better explanation	0.460*** (0.088)
Reason: Both trustworthy	0.287 (0.429)
Reason: Both unstrustworthy	0.157 (0.234)
Reason: Citation	0.470*** (0.093)
Reason: First sentence	0.453** (0.211)
Reason: More reasonable	0.027 (0.154)
Reason: More technical	−0.108 (0.072)
Constant	0.002 (0.273)
Observations	345
R^2	0.256
Adjusted R^2	0.164
Residual Std. Error	0.457 (df = 306)
F Statistic	2.772*** (df = 38; 306)
Note:	*p<0.1; **p<0.05; ***p<0.01


Why Treatment Respondents Choose Article B.
count
30
20
10
0
Why?
Prefer not to answer
better explanation
both trustworthy
both untrustworthy
citation
first sentence
more reasonable
more technical


Why Treatment Respondents Choose Article A.
count
40
30
20
10
0
Why?
Prefer not to answer
better explanation
both untrustworthy
citation
more reasonable
more technical


Why Control Respondents Choose Article B.
count
30
20
10
0
Why?
Prefer not to answer
better explanation
both trustworthy
both untrustworthy
first sentence
more reasonable
more technical


Why Control Respondents Choose Article A.
count
40
20
0
Why?
Prefer not to answer
better explanation
both trustworthy
both untrustworthy
citation
first sentence
more reasonable
more technical