Does Clickbait Lead to More Clicks?

Hsuanyi Cheng, Patrick Kim, Kayla Wopschall, Ziwei Zhao

8/12/2021

Abstract

We study the effect on click-through rates of applying textual and stylistic image often related to clickbait to headlines of newspaper articles which can be bought in a digital environment. Publishing a dataset consisting of original headline, rewritten headline and re-applying image in social media and survey platform, we can directly measure whether these "clickbait features" do what they are believed to do: entice readers to click on them. The main findings are as follows. First, the data shows that image or both text and image click-bait features lead to a statistically significant increase in the number of clicks. Second, re-write headline-only articles shows no statistically significant increase in the number of clicks. The findings are concluded by two different experiments through various platforms.

Introduction

Still needs editing

The way people consume newspaper articles is changing: more and more newspaper articles are consumed on the internet rather than from physical newspapers. People used to buy a newspaper, read it from cover to cover while scanning headlines, and reading articles that they thought were interesting (Holmqvist et al. 2003). However, increasingly more people are reading individual articles online, outside of their original publication. Often, a person reads this article because it was shared on social media or some other internet platform. According to Journalism portal, in a 2017 survey, they found two-thirds of readers utilized social media to get fresh news.

With this change, the function of the headline of a news article changed as well. Previously, the primary function of a headline and image were to give the reader, who was scanning the newspaper, a clear understanding of what the article was about. But in social media the headline and image become primary ways to attract the reader's attention and make the reader curious as to what the article is about, so that it lures the reader to open the article. Thus we started to wonder if attractive headlines and images, often called click-bait, did draw more attention than regular newspaper headlines.

Clickbait is a commonly explored device, with varied results (Mini, 2021). While clickbait, strictly defined, is the crafting of sensational headlines and images that lead to misleading content, we want to test the impact of the clickbait strategies within various other contexts(e.g. news, politics, popular science). As content developers, you're constantly trying to drive engagement with readers, and structuring your meta content (headlines, subheadings, and images) on social media platforms should influence active engagement with your content, and the types of individuals you engage with.

In the experiment, we design the control group to view a headline that is factual, unemotional with a similarly neutral but relevant image. Often captured in the newspaper's original writing. The treatment group will be a more emotional headline and/or image that is designed to trigger someone to click through due to curiosity, anxiety, or any other feelings pushed forward by the headline. Both groups would arrive at the same article if they click through on the ad. The mechanisms we're trying to test here are people's interest is piqued more by one type of image and/or rhetoric over another, enough to engage with the content (clicks, likes, comments). Through our various treatment groups we attempt to see the effects of images, headlines, and the combination of the two.

We initially utilized Facebook ads platform to execute an A/B test, in which control and treatment image and headline will render to different users separately. In just a few days, we successfully collected more than thousand data points to analyze the click bait effects. Unfortunately, Facebook ads algorithms detected our experiments and considered it against their policy. Even though we tried to explain and provided personal identification to prove legit execution. In the end, we decided to run a survey via Survey Monkey to continue the same experiment with a different design approach. In the second stage of the experiment, we send out the survey displaying both control and treatment with different articles to avoid obvious click-bait tests. In both experiments, we are able to archive randomized and unbiased results by the platform setting.

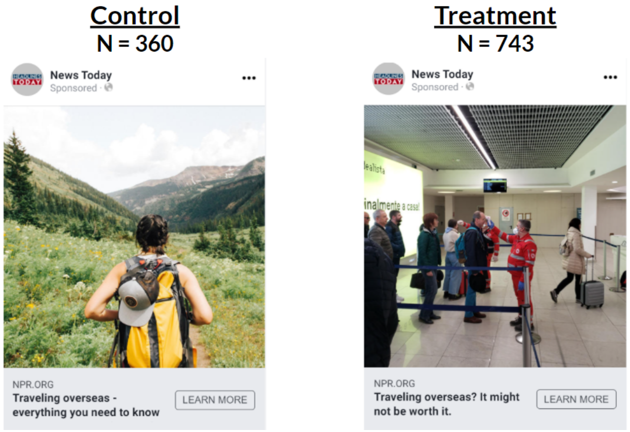
Facebook Ads Experiment

Experimental Design

We chose Facebook advertising as our preferred platform for the experiment for several reasons. First, it is one of the most widely used social media platforms, which would allow flexibility in choosing the sample population and ensure that we have enough subjects in the experiment given our limited budget. Secondly, Facebook has an intuitive A/B testing feature that allows us to easily design ad campaigns with a control ad and a treatment ad that never get shown to the same Facebook user.

We conducted two pilot experiments. In each experiment, we use one existing news article and design two ads for it, each with a headline and an image. For the ad that gets shown to the control group, both the headline and image are rather neutral. They state facts without a call to action or question for the reader, and the image should not elicit strong emotional reactions. For the ad shown to the treatment group, both the headline and image are "clickbait" - the headline contains a question that entices the reader to click on the article to find the answer, and the image portray more dynamic action that might draw attention.

Figure 1 shows the two ads used for the first pilot experiment. It uses the same NPR article about traveling overseas during the pandemic. The control group sees the headline "Traveling overseas - everything you need to know" with an image of a woman hiking in an idyllic nature scene. The treatment group sees the headline "Traveling overseas? It might not be worth it" with an image of travelers getting their temperatures checked in an airport security line.

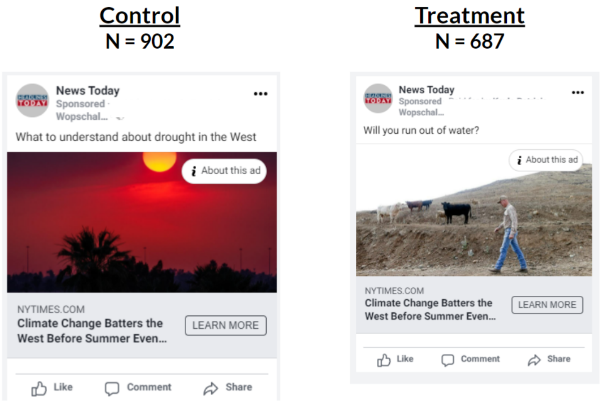


NPR.ORG
Traveling overseas - everything you need to know

NPR.ORG
Traveling overseas? It might not be worth it.

Figure 1: Facebook Pilot Experiment 1

Figure 2 shows the two ads used for the second pilot experiment. It uses an New York Times article about the drought on the West Coast. The control group sees the headline "What to understand about the drought in the west” with an image of a red fiery sky. The treatment group sees the headline “Will you run out of water?” with a farmer walking on barren land in front of a few cows.



News Today
Sponsored
Wopschal...

What to understand about drought in the West

NYTIMES.COM
Climate Change Batters the West Before Summer Even...

News Today
Sponsored
Wopschal...

Will you run out of water?

NYTIMES.COM
Climate Change Batters the West Before Summer Even...

Figure 2: Facebook Pilot Experiment 2

Each experiment got to run for about two days. Because the business page from which we created ads is brand new, and we did not have much, if any, previously advertising records on Facebook, our respective advertising accounts got disabled by Facebook. It was eye opening to see efforts being made by social media platforms to limit “fake news”, even though it was a big roadblock for our experiment. However, we were able to extract data from these pilot studies and see some meaningful results.

Results

Regression Model

The first regression analysis is linear regression model of the variable click on the variable treated - both indicator variables. Click indicates whether a Facebook user clicked through to the article after seeing the ad, and treated indicates if a user is in the treatment group. Table 1 shows the results. The first column shows the results from the first pilot study using the travel ads. The coefficient on treated is 0.122 with a 0.00 p-value, meaning that people in the treatment group are 12.2% more likely to click on the ad. The second column shows results from the second pilot study using the drought ads. The coefficient on treated is 0.020 with a p-value of 0.0691. People in the treatment group are 2.0% more likely to click on the ad, but it is only only statistically significant at the 10% level, compared to the 1% level in the first pilot study.

Table 1: Facebook experiment - click regressed on treated

	Dependent variable:	
	click	
	travel ads	drought ads
	(1)	(2)
treated	0.122***	0.020*
	(0.016)	(0.011)
Constant	0.031***	0.040***
	(0.009)	(0.007)
Observations	1,103	1,589
R^2	0.033	0.002
Adjusted R^2	0.032	0.001
Residual Std. Error	0.311 (df = 1101)	0.215 (df = 1587)
F Statistic	37.035*** (df = 1; 1101)	3.308* (df = 1; 1587)
Note:		*p<0.1; **p<0.05; ***p<0.01

Covariate Check Since Facebook's algorithms randomized our populations for the experiments and their business goals are presumed to be focused on revenue from advertisers, we conducted covariate checks to ensure that randomization was done properly. We conducted a regression of treated on female to see if gender predicted which group a user would fall in. The female coefficient is not statisticaly significant in the travel ads experiment, but it is significant at the 10% level in the drought ads experiment.

Table 2: Facebook experiment - covariate check - gender

	Dependent variable:	
	treated	
	travel ads	drought ads
	(1)	(2)
female	0.037	−0.042*
	(0.029)	(0.025)
Constant	0.650***	0.453***
	(0.024)	(0.018)
Observations	1,103	1,589
R^2	0.001	0.002
Adjusted R^2	0.001	0.001
Residual Std. Error	0.469 (df = 1101)	0.495 (df = 1587)
F Statistic	1.600 (df = 1; 1101)	2.897* (df = 1; 1587)
Note:		*p<0.1; **p<0.05; ***p<0.01

We ran a similar regression of treated on age. Age was provided from facebook in buckets: 18-24, 25-34, 35-44, 45-54, 55-64 and 65+. The regression results effectively compare each age group to the 18-24 group. For both the travel ads experiment and the drought ads experiment, all coefficients on age groups are positive. None of the coefficients are statistically significant in the travel ads experiment, but they are all statistically significant at different levels for the drought experiment. Without other covariates or a deeper understanding of Facebook's algorithm, it is difficult for us to tell why these the drought experiment's randomization may not have been done properly.

Table 3: Facebook experiment - covariate check - age

	Dependent variable:	
	treated	
	travel ads	drought ads
	(1)	(2)
as.factor(age)25-34	0.260	0.061**
	(0.713)	(0.027)
as.factor(age)35-44	0.297	0.227**
	(0.709)	(0.091)
as.factor(age)45-54	0.182	0.249*
	(0.708)	(0.147)
as.factor(age)55-64	0.197	0.526***
	(0.708)	(0.055)
as.factor(age)65+	0.126	0.570***
	(0.707)	(0.032)
Constant	0.500	0.366***
	(0.707)	(0.016)
Observations	1,103	1,589
R^2	0.010	0.085
Adjusted R^2	0.006	0.082
Residual Std. Error	0.468 (df = 1097)	0.475 (df = 1583)
F Statistic	2.279** (df = 5; 1097)	29.540*** (df = 5; 1583)

Note: *p<0.1; **p<0.05; ***p<0.01

Survey(Monkey) Experiment

Because we spent so much time trying to do the Facebook experiment, we unfortunately had to redesign, deploy, and analyze a new experiment shortly before the deadline. We chose to utilize SurveyMonkey because their platform allows for surveying users cheaply and efficiently with little danger of being locked out of the platform. The respondent is shown two image ads and asked to indicate which ad they would be more likely to click on. The survey targeted 400 respondents from across the United States using the broadest demographic available on SurveyMonkey. The full dataset contains 437 responses as SurveyMonkey provided these extra responses for free.

Experimental Design

SurveyMonkey contains an A/B testing feature that allows for different text and images to be shown to respondents at random. We utilize this feature to show two different images to the respondent and create many sampling pairs. First, two ads with tame language and normal stock images were created by the team. For each ad, one version is created with extreme language in the title, another version with a shocking image, and a final version with both modified at once. This means that there are four versions of each article, and each version has a 25% chance of being displayed to the respondent according to SurveyMonkey's randomization procedure.

Here it is important to note a critical difference between the survey design and the Facebook Ads experiment. In the Facebook Ads experiment it was an organic ads placement where an individual had the option to click or ignore. We were able to capture the data for those who observed and chose not to click. In our survey design, the survey respondents are forced to give a selection on their preference. This is critical for several reasons: 1) it requires a binary choice which doesn't reflect a normal news engagement environment; 2) it places the preferences of those who are uninterested in both as equal to those who have a strong preference to one, adding noise from individuals that may not have clicked on either in a real world scenario; 3) as a survey the respondent doesn't actually get to engage in the content, where as in Facebook these were sitting over a real article with content - making the entire format of engagement very different.

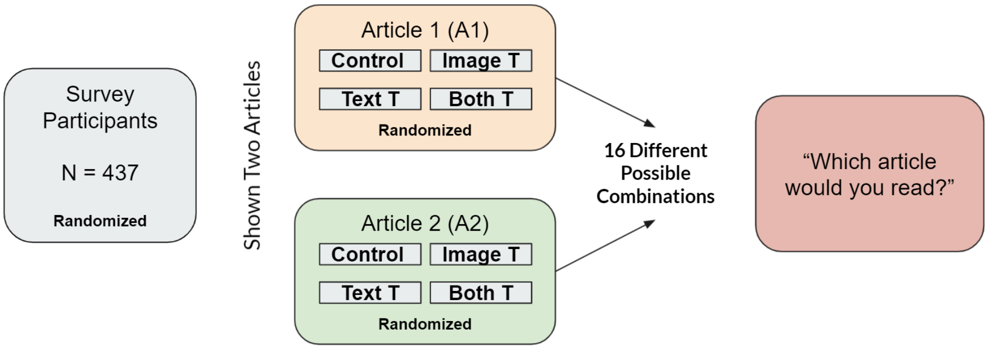


Figure 3: Survey Design

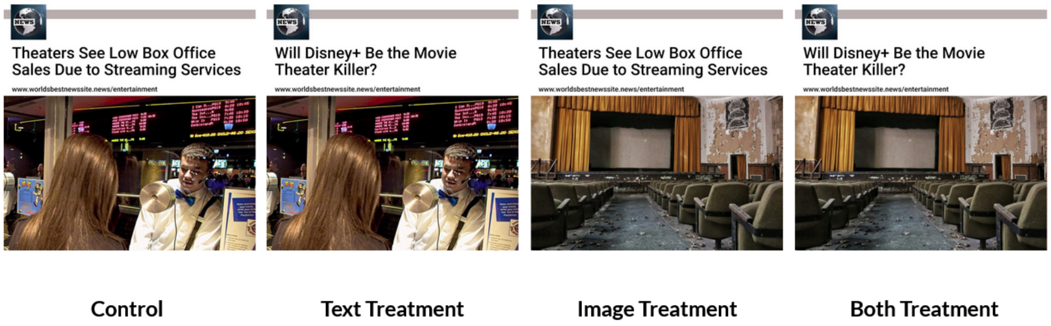

NEWS
Theaters See Low Box Office Sales Due to Streaming Services
www.worldsbestnewssite.news/entertainment
Will Disney+ Be the Movie Theater Killer?
www.worldsbestnewssite.news/entertainment
Theaters See Low Box Office Sales Due to Streaming Services
www.worldsbestnewssite.news/entertainment
Will Disney+ Be the Movie Theater Killer?
www.worldsbestnewssite.news/entertainment
Control
Text Treatment
Image Treatment
Both Treatment

Figure 4: Article 1

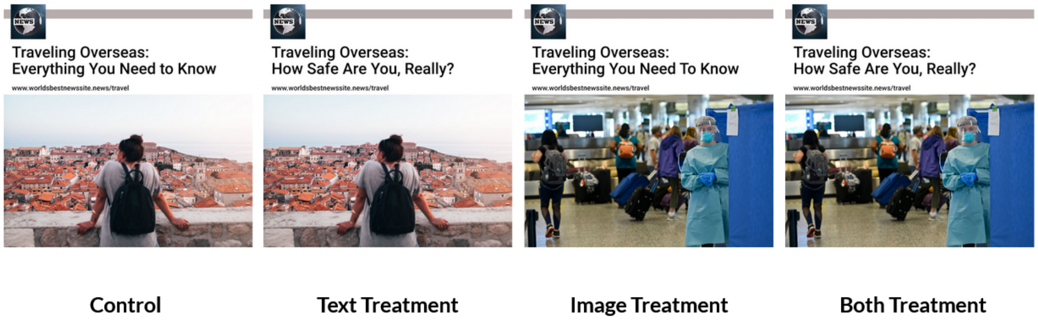
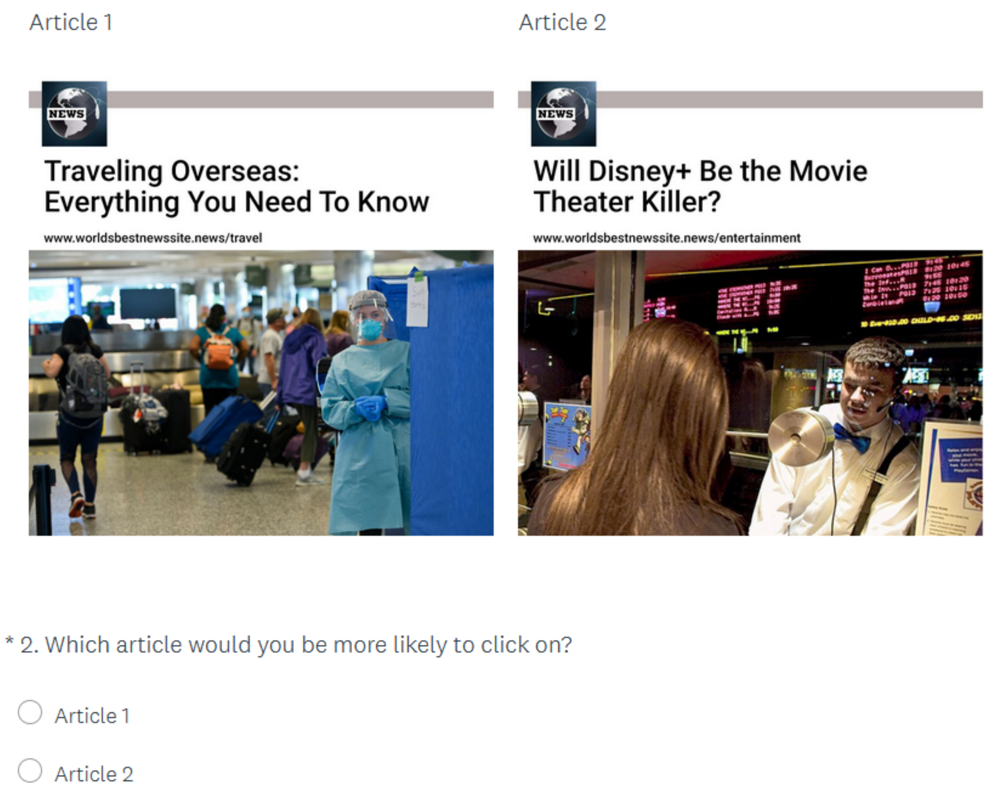

NEWS
Traveling Overseas: Everything You Need to Know
www.worldsbestnewssite.news/travel
Traveling Overseas: How Safe Are You, Really?
www.worldsbestnewssite.news/travel
Traveling Overseas: Everything You Need To Know
www.worldsbestnewssite.news/travel
Traveling Overseas: How Safe Are You, Really?
www.worldsbestnewssite.news/travel
Control
Text Treatment
Image Treatment
Both Treatment

Figure 5: Article 2

Since there are two ads and four possible versions of each ad, this creates 16 possible pairings, meaning that each pairing has a 1 in 16 chance of appearing in a respondent's survey.



Article 1

Traveling Overseas: Everything You Need To Know

www.worldsbestnewssite.news/travel

Article 2

Will Disney+ Be the Movie Theater Killer?

www.worldsbestnewssite.news/entertainment

* 2. Which article would you be more likely to click on?

○ Article 1

○ Article 2

Figure 6: Survey Sample

Results

Respondents only ever see one version of each ad, which means that the potential outcomes of a respondent observing a treated ad and a control ad are never observed. To overcome this problem, a test of proportions is used as a basic check to determine whether the difference in ad content or topics has a significant effect on respondent choices. For each treatment type, we compare the number of times that ad version was selected by the respondents between both ads, and the p-values obtained are outlined below.

Test of Proportion P-values:

- Control group - those who saw control images of both articles: 0.6510766
- Text treatment group - those who saw text treatment of both articles: 1
- Image treatment group - those who saw image treatment for both articles: 0.3319755
- Both treatment group - those who saw both text and image treatment on both articles: 0.4414183

For all groups we see no statistically significant differences in preference for article one versus article two, suggesting content and topic are not driving the outcome.

Test of Article Preference Through Regression Analysis: Due to small sample sizes, we also ran a regression on a subset of the dataset containing all of the groups outlined above (e.g. those who saw the same type of both articles). Here we have the outcome variable set as preferring article one, and covariates for potential treatment types. We see, even when aggregated, there is no evidence of individuals in the survey preferentially selecting one article over the other when both articles have equal levels of treatment. Therefore, there is no evidence that article content is driving article selection, and we can assume that effects seen are driven by differences in treatment.

Table 4: Survey Monkey experiment - article content preference check

	Dependent variable:
	article_pref groups
treatmentC	−0.138 (0.125)
treatmentI	−0.240 (0.158)
treatmentT	−0.069 (0.151)
Constant	0.593*** (0.098)
Observations	109
R^2	0.025
Adjusted R^2	−0.003
Residual Std. Error	0.503 (df = 105)
F Statistic	0.898 (df = 3; 105)
Note:	*p<0.1; **p<0.05; ***p<0.01

Treatment Type Impacts at the Individual Level

To test the impact of treatment types at the individual level, we subset the data to remove all individuals that saw identical treatment levels (e.g. control article one and control article two; treatment both article one and two, etc). This left us with individuals that all have one control article, and see one type of treatment article, then make the choice between the two for their preference.

When running the regression, we then have a dummy variable that notes whether article one is control or treatment, and variables for the types of treatments that were seen (Image, Text, or Both). Our outcome variable is whether the treatment article was selected or not.

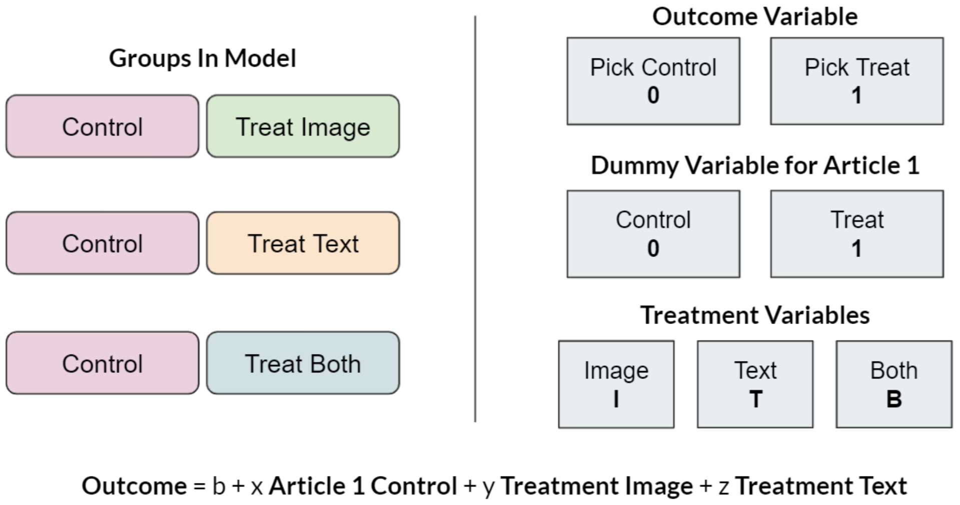


Figure 7: Individual Treatement Design

Table 5: Survey Monkey experiment - regression of covariates

	Dependent variable:
	outcome
article1_control	0.043 (0.079)
treatment_typeI	-0.310^{***} (0.088)
treatment_typeT	-0.237^{**} (0.100)
Constant	0.629^{***} (0.076)
Observations	173
R^2	0.073
Adjusted R^2	0.057
Residual Std. Error	0.486 (df = 169)
F Statistic	4.445^{***} (df = 3; 169)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression Analyses and Results From these results we see that when Article One is the control, respondents that are shown an image with "Both" image and text are 67% more likely to select the treatment article over the control article. When the treatment variable is just image, the percentages decrease by 31%, suggesting that when looking at Article One control and Article Two with a treatment image they are 36% more likely to select the treatment article. If the treatment applied is "Text", the overall likelihood of selecting the treatment article is 45%. When article two is made the control, all of these percentages decrease by 4.3%.

Overall, this suggests significant impacts of all types of treatments. The largest treatment impact belonging to the cases where both image and text are changed, followed by text being the second most impactful, and images being the third most impactful.

Treatment Type Impacts at the Group Level

While it appears that these variables correlate on an individual level with article selection, our next question was focused around if we see any statistical significance when compared to a control group. To answer this question we ran four different models with the following comparisons as control and treatment:

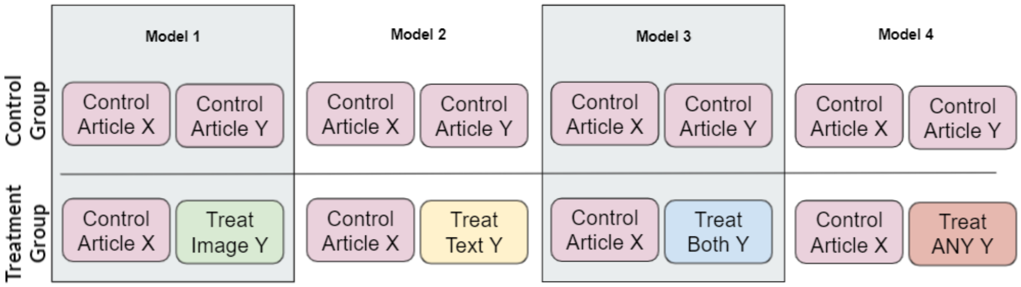
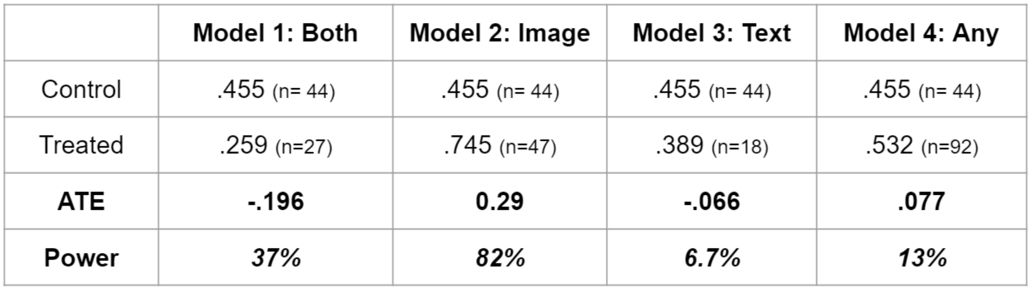


Figure 8: Control and Treatment Groups

The treatment group that includes "ANY" treatment is an aggregation of the prior three models/groups. This was run twice for article one and two - where each article is independently run for being the control.

Article One as Control Keeping article one as control we see an average treatment effect of the following for each comparison:



	Model 1: Both	**Model 2: Image**	**Model 3: Text**	**Model 4: Any**
Control	.455 (n= 44)	.455 (n= 44)	.455 (n= 44)	.455 (n= 44)
Treated	.259 (n=27)	.745 (n=47)	.389 (n=18)	.532 (n=92)
ATE	**-.196**	**0.29**	**-.066**	**.077**
Power	***37%***	***82%***	***6.7%***	***13%***

Figure 9: Article One as Control: Summary Statistics

When regressing for these variables where article one is always the control we see the following results:

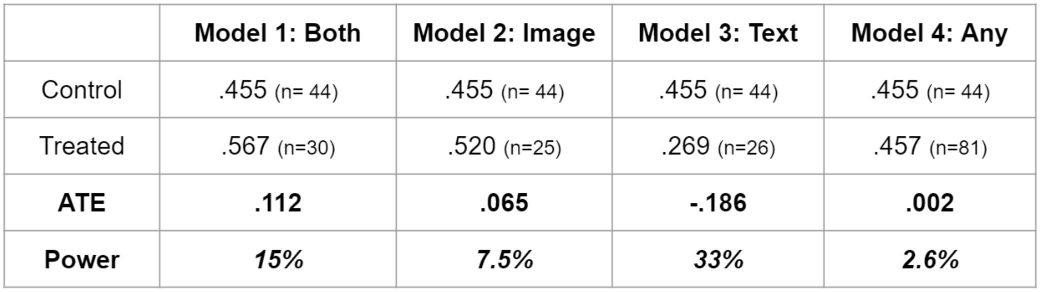
With Article One as our control, we can see that there appears to have a significant effect when looking at the Image treatment of .290, and a negative effect when looking at comparisons with "Both" image and text treatment of -.195. However, when looking at our summary statistics and power of these tests we can note that by parsing our data in this manner we are now working with small samples sizes and low power. Interestingly, our one variable that shows a positive significant positive correlation in our regression, the articles with image treatment, is the only group comparison that has a high power associated.

From these results, we can lend more evidence that there may be some impact related to these treatment effects, however with this particular analyses we suffer a sample size issue to better articulate the nuance of those impacts.

Table 6: Survey Monkey experiment - Group Comparisons

	Dependent variable:			
	article_pref			
	Both	Image	Text	Any Treatment
	(1)	(2)	(3)	(4)
treated	−0.195*	0.290***	−0.066	0.078
	(0.116)	(0.101)	(0.144)	(0.093)
Constant	0.455***	0.455***	0.455***	0.455***
	(0.077)	(0.077)	(0.077)	(0.077)
Observations	71	91	62	136
R^2	0.038	0.088	0.004	0.005
Adjusted R^2	0.024	0.078	−0.013	−0.002
Residual Std. Error	0.483 (df = 69)	0.472 (df = 89)	0.503 (df = 60)	0.502 (df = 134)
F Statistic	2.736 (df = 1; 69)	8.579*** (df = 1; 89)	0.218 (df = 1; 60)	0.719 (df = 1; 134)
Note:				*p<0.1; **p<0.05; ***p<0.01

Article Two as Control Keeping article two as control we see an average treatment effect of the following for each comparison:



	Model 1: Both	Model 2: Image	Model 3: Text	Model 4: Any
Control	.455 (n= 44)	.455 (n= 44)	.455 (n= 44)	.455 (n= 44)
Treated	.567 (n=30)	.520 (n=25)	.269 (n=26)	.457 (n=81)
ATE	**.112**	**.065**	**-.186**	**.002**
Power	***15%***	***7.5%***	***33%***	***2.6%***

Figure 10: Article Two as Control: Summary Statistics

When regressing for these variables where article two is always the control we see the following results:

With Article Two as our control, we see no statistically significant coefficients with our regressions, however we also see from the summary table that we are suffering from a power issue related to our sample sizes.

At this time its best to use these group comparisons as evidence of a potential impact, seeing as our one comparison with power shows a signficiant coefficient, however further work to grow our sample size needs to be pursued before anything concrete can be concluded.

Conclusions

Still needs editing

In the Facebook ads platform experiment, the results suggest that with 'clickbait' images and text get significantly more clicks. In the survey results, we analyzed as "within subjects" and "between subjects". In the Within subjects analysis, it suggests that both Image and text modifications have a positive ATE, and

Table 7: Survey Monkey experiment - Group Comparisons

	Dependent variable:			
	choose_article1			
	Both	Image	Text	Any Treatment
	(1)	(2)	(3)	(4)
treated	−0.112	−0.065	0.185	−0.002
	(0.121)	(0.129)	(0.119)	(0.095)
Constant	0.545^{***}	0.545^{***}	0.545^{***}	0.545^{***}
	(0.077)	(0.077)	(0.077)	(0.077)
Observations	74	69	70	125
R^2	0.012	0.004	0.034	0.00000
Adjusted R^2	−0.002	−0.011	0.020	−0.008
Residual Std. Error	0.504 (df = 72)	0.506 (df = 67)	0.485 (df = 68)	0.502 (df = 123)
F Statistic	0.883 (df = 1; 72)	0.267 (df = 1; 67)	2.382 (df = 1; 68)	0.001 (df = 1; 123)
Note:				*p<0.1; **p<0.05; ***p<0.01

lead to the most preferences for 'clickbait' type articles. Image and Text modifications both have significant positive ATE's, with Text being slightly more impactful than image.

In the Between subjects analysis, it suggests that when Article 1 is Control, there is a positive significant ATE from Image Treatment and a negative significant ATE from Both treatments combined. But when Article 2 is controlled, there is no significant ATE from Article 1 Treatments.

Overall, we believe that sensational headlines and images did attract people's attention and the image has a more dramatic effect than the text. Our experiment's result can only reveal partial evidence due to unexpected platform difficulties. To extend this experiment and analysis, we could navigate different platforms and various content types.