The Effect of “Natural” in Marketing Claims on Consumers’ Interests in Online Advertising

Fall 2021 W241 Final Project Report
Group: Amber Chen, Chris Skokowski, Linh Nguyen, Sissie Cui

Introduction

The term “natural” is frequently used in the marketing claims for food, baby, skin care, and many other types of products. However, unlike terms such as “organic” or “non-GMO,” the use of “natural” product labelling is not regulated by any government authorities[1] and uncertified marketing language could impact how customers perceive a product. We investigate the effects of this language on consumers’ interests by running experiments to see differences in online ad click-through rates when using such uncertified marketing claims in product images and descriptions.

This work is driven by growing consumer awareness of green product features. To meet consumers’ interests in healthy and green products, retailers and producers put terms and labels such as “natural”, “non-GMO” and “organic” on product packaging and marketing claims. The term “natural” topped the list of green product features that resonate best with consumers. However, unlike “non-GMO” and “organic”, the use of “natural” is often not regulated by government authorities. Nonetheless, some research publications show that consumers seek out “Natural” labelled products even though such labeling is unregulated. Based on the study conducted by Euromonitor International, 55% of survey respondents look for natural features when buying products in at least one category, compared with 41% who look for eco-friendly and 39% who look for organic. Some respondents also indicated that they are willing to pay at a premium price to purchase products with a “natural” label or marketing claim.[2]

The goal of our study is to understand whether uncertified marketing language impacts consumers’ perception of a product. In our experiment, we have investigated the research question below: *Do consumers have more interest in products when the term “natural” is included in marketing claims?* We have used statistical hypothesis testing to evaluate the null hypothesis that including the term “natural” in marketing claims has no treatment effect on consumers’ interests. The alternative hypothesis is that including the term “natural” marketing claims has a non-zero treatment effect on consumers’ interests.

Experiment Overview

We used Facebook Ads Center to deliver advertising campaigns for fictitious products with differences in language and labeling across treatment groups. A 2x3 factorial design was utilized to investigate the treatment effects across 3 different types of products: produce, clothing, and auto care products. Facebook users in the treatment groups received an advertisement that boasts the use of “all natural” ingredients or materials in the description, as well as a “100% natural” label on the advert images. Facebook users in the control groups received the same ads without the “all natural” language in the description and the “100% natural” label on the images. Screenshots of the advertisements that were delivered to users are provided in Appendix A.

The Facebook Ads Center performance metrics provided us with results in terms of clicks and impressions, in addition to age group and gender (the primary demographic covariates of interest). From the provided analytics reports, we calculated the clickthrough rate for the advertisements (CTR = clicks / impressions), which served as the outcome of interest. Therefore, the treatment effects for this study were the differences in CTR between the treatment and control groups. This represents the differences in the potential outcomes (clickthrough rate) between users shown ads with “natural” language and the potential outcomes for users shown ads without “natural” language.

Users who clicked on the advertisements were directed to a minimal “Coming Soon” landing page for a fictitious company. Since our outcome of interest is CTR rather than product sales, this method suited our experimental design and simplified the logistics of implementing the experiment. Examples of landing pages are provided in Appendix B.

Power Calculation

Prior to initiating the advertising campaigns, we conducted a power analysis to determine the minimum sample sizes required to detect treatment effects. The power calculation was conducted using the following equation:

$$\beta \ = \ \Phi\left(\frac{\left|\mu_t - \mu_c\right|\sqrt{N}}{2\sigma} - \Phi^{-1}(1 \ - \ \alpha/2)\right)$$

$\boldsymbol{\beta}$ represents the statistical power of the experiment, which we set at or above 0.80 by convention. α represents desired statistical significance (0.05 by convention). $\boldsymbol{\Phi}$ is the CDF and inverse CDF of the standard normal distribution. The two values of μ represent the average clickthrough rates in the treatment and control groups. Finally, N and σ represent the sample size and overall standard deviation of the outcome variable.

We can examine the data on click-through rates (CTRs) for facebook ads to estimate μ and σ. For even greater precision, we can adjust for the difference in average click-through rate by industry to account for expected differences across our three sets of ad campaigns. For clothing and auto care products, a recent report[3] found that Facebook ads had an average CTR of 0.0111 in 2021, with CTRs of 0.0124 and 0.0080 for apparel and auto ads, respectively. For data on food advertisements, another report[4] using data from 2019 reported an overall average CTR of 0.0089 and a CTR of 0.0120 for food and drink. Since the overall average CTR increased between 2019 and 2021, we scale up the food CTR by the ratio of overall average CTRs, which gives us an adjusted CTR of 0.0150 for food products.

To estimate the difference between treatment and control means, we select a conservative estimate of a 30% positive treatment effect (far below the 55% effect observed in previous literature referenced in the Introduction section). By using the power calculation equation and confirming results with a sample size calculator[5], we obtained sample size requirement estimates for all 3 product types, provided below in Table 1.

Product Type	β	α	μ	$\left\|\mu_t - \mu_c\right\|$	**N**
Apparel	0.8	0.05	0.0124	0.00372	31,888
Auto Care	0.8	0.05	0.0080	0.00240	49,682
Produce	0.8	0.05	0.0150	0.00450	26,280

Table 1: Parameters for statistical power calculation

Based on these findings, we ran our ad campaigns and allocated budgets for each advertising campaign accordingly. We generally aimed to reach between 25,000 to 50,000 impressions for each ad campaign by defining our advertising budget within Facebook Ads Center.

Flow Diagram of Observations

Our flow diagram of observations throughout the experiment (Figure 1) demonstrates the number of ad impressions for each treatment group and product type in our study. Since we obtained fairly high click through rates (as demonstrated in the Analysis & Results section) and had a large amount of ad impressions for each treatment group, we were able to achieve high statistical power for this study and obtain statistically significant results.

It is infeasible to monitor noncompliance in our study (e.g., Facebook users looking away from their screen as they scroll past one of our advertisements). We therefore concern ourselves with estimating the average treatment effect for the entire population (rather than the complier average causal effect) and assume that noncompliance rates are similar across the treatment and control groups. As such, no noncompliance or attrition is reflected in our flow diagram of observations below (Figure 1).

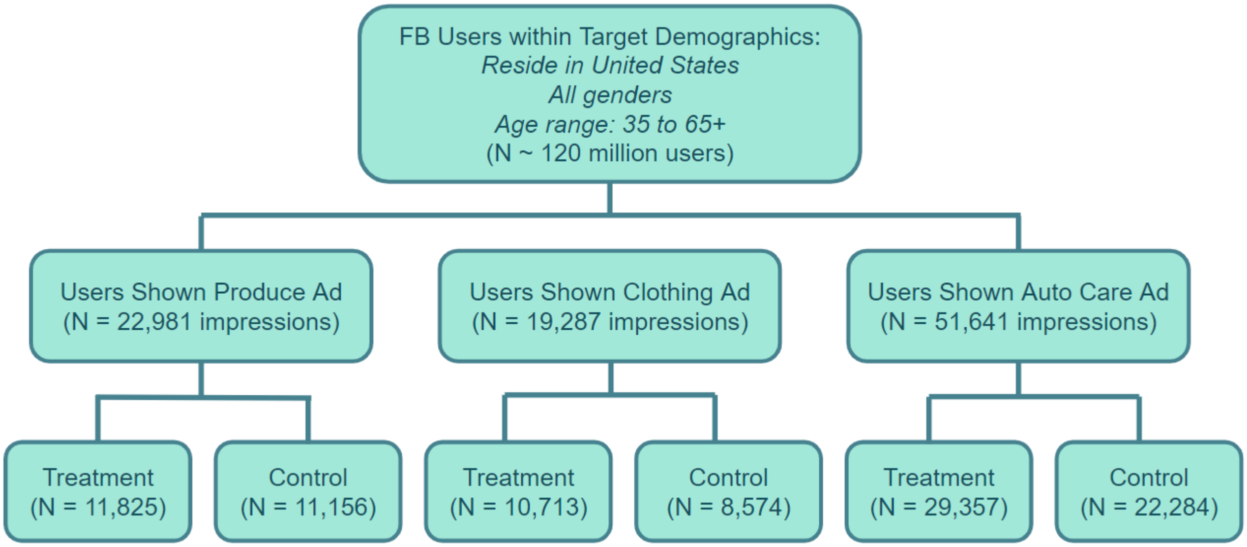


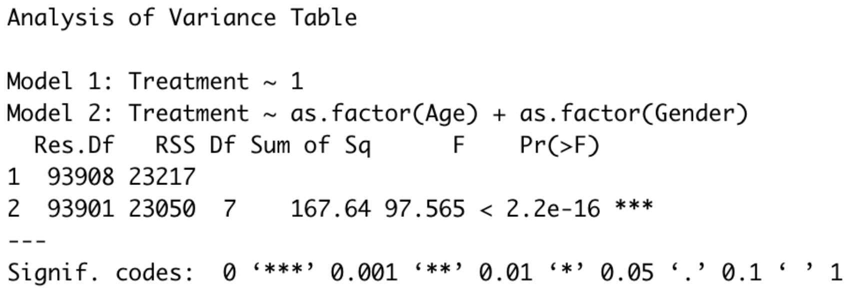
Figure 1: Flow diagram of observations throughout the experiment

Randomization

The random assignment of users to be shown either a treatment or control advertisement was conducted by Facebook's A/B testing platform. For all our advertising campaigns, we defined the same set of broad target demographics (Facebook users residing in the US, all genders, age range 35 to 65+). The A/B testing platform ensures that no users who are shown the treatment advertisements are also shown the control advertisements, and vice versa. We defined a fairly broad target audience in the interest of being able to generalize our results as much as possible.

A disadvantage of using Facebook Ads Center to run the advertising campaigns is that we did not have direct control of the randomization process. Although we defined the same sets of target demographics for every ad campaign, Facebook still conducts some background selection of users to receive our advertisements that we had no control over. The large differences in impressions between product types (~23k impressions for produce ads, ~19k impressions for clothing ads, ~52k impressions for auto care ads) can be attributed to this background selection work. The differences between the number of impressions across treatment and control groups for each advertising campaign are still quite small (e.g., 11,825 treatment impressions and 11,156 control impressions for produce ads). Still, we need to assess whether there are systematic differences between the users who are shown advertisements with "natural" language by Facebook's selection algorithm and those who are shown the control advertisements.

In order to assess the randomization of our treatment and control groups, we use an F-test with the null hypothesis that the distributions of covariates within each group are identical. To conduct this test, we create one linear model that uses only a single constant as a feature in predicting treatment, and another linear model that uses the covariates of the population (age and gender) as features for predicting treatment. If the covariate model explains more variance in the treatment variable than the constant model does, then the treatment and control populations do not have the same distribution of covariates, meaning they were not properly randomized. Conducting an F-test of this format on the entirety of our data gives us the following result:



```
Analysis of Variance Table

Model 1: Treatment ~ 1
Model 2: Treatment ~ as.factor(Age) + as.factor(Gender)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1  93908 23217
2  93901 23050  7    167.64 97.565 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: ANOVA table for randomization check F-test

These results clearly indicate a difference in covariate distributions between the treatment and control groups overall, meaning we reject the null hypothesis that the treatment and control groups share the same distributions of covariates. In order to determine whether this was a problem specific to certain campaigns, we ran the same test on each subset of data associated with a specific campaign. The results are displayed in the table below:

Campaign	**F-Value**	**P-value**
Auto Care	13.504	3.43e-13
Clothing	80.369	2.2e-16
Produce	80.528	2.2e-16

Table 2: Randomization check F-test results for each campaign

As we can see, the “Auto Care” campaign had marginally better randomization than the Clothing and Produce campaigns, but none had effective randomization protocols, meaning the null hypothesis of effective randomization is rejected for each campaign.

Since we could not observe treatment and control groups before running our ad campaigns, we had little to no control over the randomization process employed by Facebook. While Facebook’s A/B testing platform was able to build treatment and control groups of roughly the same size, it was not able to produce a properly randomized partition of the two groups. This somewhat disappointing result would be a primary area for improving on our work in future research.

Data Completeness

Randomization aside, our data isn't quite as complete as we would like it to be. Despite setting our advertisements to include the maximum population possible, we only received a very limited amount of impressions from the 18-24 and 25-34 age groups. The distribution of age groups across our sample is shown in the figure below:

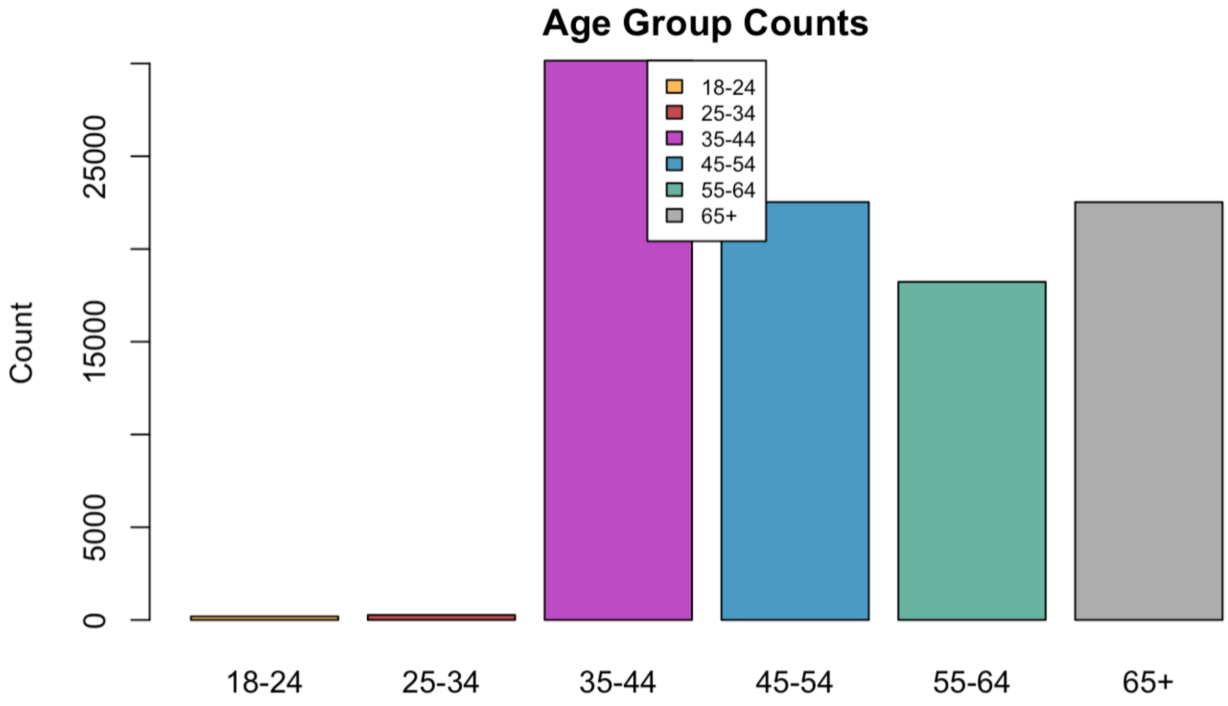


Figure 2: Distribution of Age Groups

Beyond the limited number of impressions on younger audiences, we are satisfied with the completeness of our data. Since our experiment did not involve specific instructions for treatment administration or a lengthy study, we have no need to worry about compliance among treatment groups or dropout over time. The only hiccup we had in our data collection process came when our ads were briefly taken offline by Facebook due to the fact that our storefront was not functional. This occurred on several occasions, but we cleaned data from those days with partial data from our main dataset before proceeding with our analysis. We'll proceed with a summary of our analysis of this dataset in the next section.

Data Analysis & Results

Overall Treatment Effect

Overall, we saw a negative treatment effect in the treatment group which indicated that the click-through rate was lower for the ads with "All Natural" in the description and "100% Natural" claim in the creatives. As shown in Figure 3, the click-through rate of the control group was 2.7% which was higher than the click-through rate of the treatment group, 1.69%.

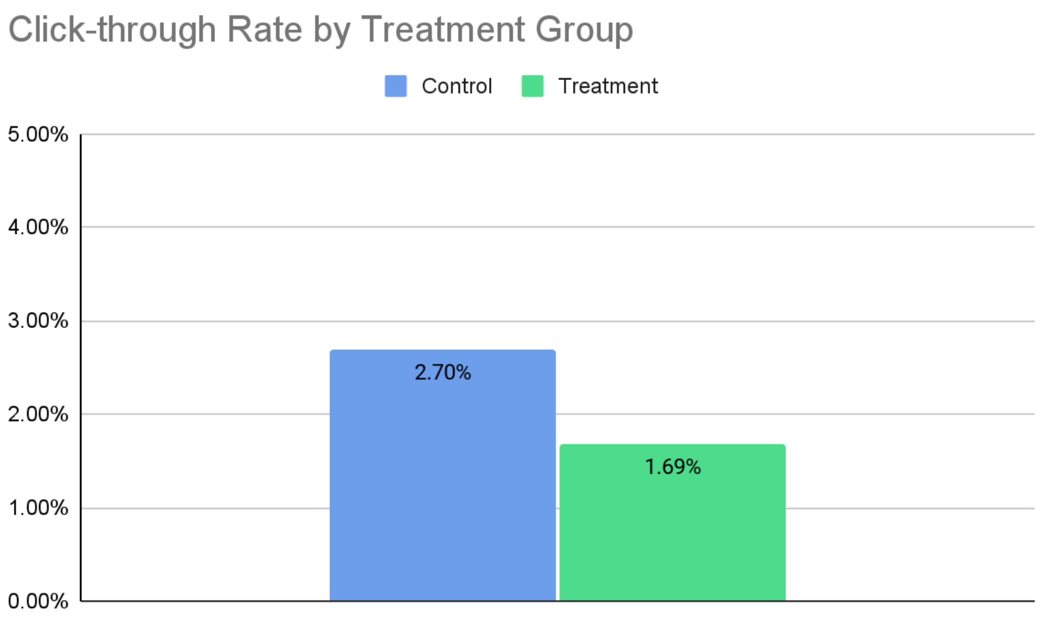


Figure 3: Click-through Rate by Treatment Group

Looking at the overall treatment effect at the product category level in Figure 4, we can see consistent negative treatment effects across the product categories. The largest difference was observed in the clothing category, and the produce category saw the smallest difference between treatment and control group.

We also developed three models which we used to determine the magnitude and statistical significance of treatment on click through rates. The first model used treatment as the only feature for predicting CTR, the second model also included age as a feature, and the final model included treatment, age, and gender as features. We performed t-tests on each of these models, and the results of those tests are displayed in the table below:

Model	Treatment coefficient	t-value	p-value
1 (Treatment)	-0.01013003	-10.423	< 2.2e-16
2 (Treatment + Age)	-0.00983305	-10.0847	< 2.2e-16
3 (Treatment + Age + Gender)	-0.00969799	-9.9362	< 2.2e-16

Table 3: T-test Results on the Three Models

We can see in the table above that the models also learned to associate treatment with an approximately 1% decrease in click through rate. Moreover, each of the models also reached that conclusion with high certainty, as evidenced by the low p-values.

We then decided to examine the results on a per-campaign basis, in order to determine whether this trend held across the different products we advertised.

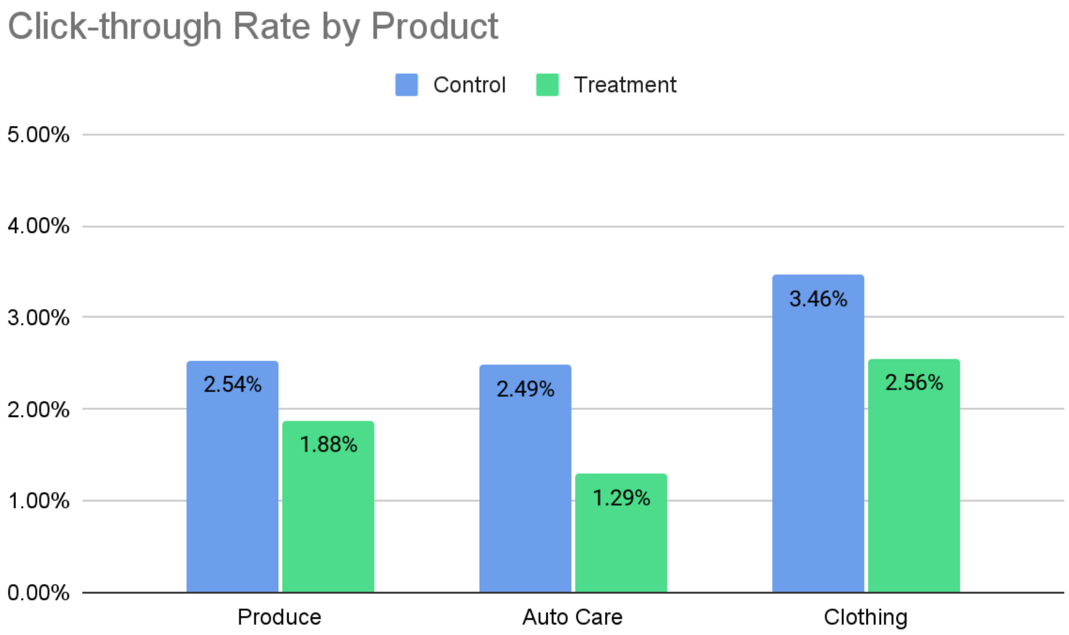


Figure 4: Click-through Rate by Product

We can see in Figure 4 that each campaign had the same general trend, with treatment being associated with lower click through rates. The overall click through rates of the campaigns differed from each other, and the differences between ctrs in control and treatment groups were also marginally different, but each campaign saw between a 0.6 and 1.2 difference between click through rates between groups.

Examining the t-test results for models trained on campaign-specific data tells us a similar story:

	Coefficient (p-value)		
Model	Produce	Auto Care	Clothing
1 (Treatment)	**-0.0065937** (0.0006904)	**-0.0119508** (< 2.2e-16)	**-0.0090632** (0.0002823)
2 (Treatment + Age)	**-0.0070032** (0.0003784)	**-0.01188844** (< 2.2e-16)	**-0.0044427** (0.0777045)
3 (Treatment + Age + Gender)	**-0.00697197** (0.0004021)	**-0.01185829** (< 2.2e-16)	**-0.00389270** (0.123105)

Table 4: Treatment Coefficients by Product on the Three Models

As we can see in the table above, each model learned to associate treatment with a reduction in click-through rate for every product. However, unlike in the previous section, which attributed statistical significance to every learned coefficient, we can see that the treatment coefficients learned for the clothing campaign-specific data were not statistically significant to a p-value of 0.05 when covariates were involved in a model. The fact that models 2 and 3 did not find the learned coefficient for treatment in the clothing campaign to be statistically significant suggests that those covariates had a much greater impact on click through rate in that campaign than in others. In order to understand this trend, we turn to analysis on the relationship between the covariates in our dataset and the outcome variable of click through rate.

Treatment Effect by Gender

In Figure 5, we can see that negative treatment effects were observed among both male and female audiences. Male group saw a more severe negative treatment effect and the click-through rate of the treatment group was only half of that of the control group.

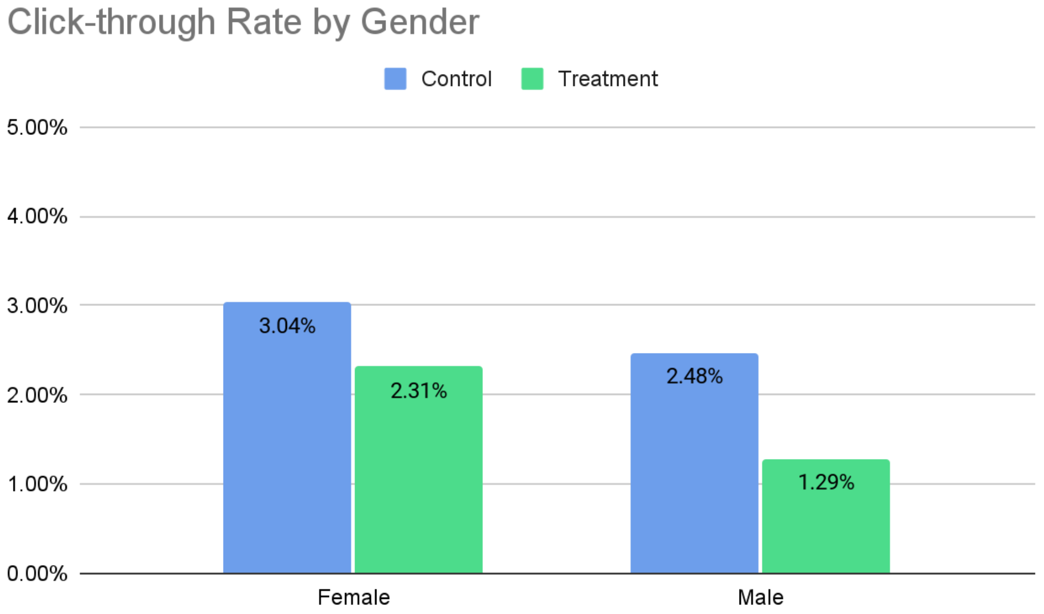


Figure 5: Click-through Rate by Gender

Unsurprisingly, female and male audiences showed stronger interest towards different categories. Overall, female audiences clicked the clothing ad the most although the treatment group still saw lower click-through rate compared to the control group. Among male audiences, the click-through rates of different products were very similar with auto care being slightly higher than the other two product categories. In addition, showing the word “natural” negatively impacted the click-through rates among both female and male audiences.

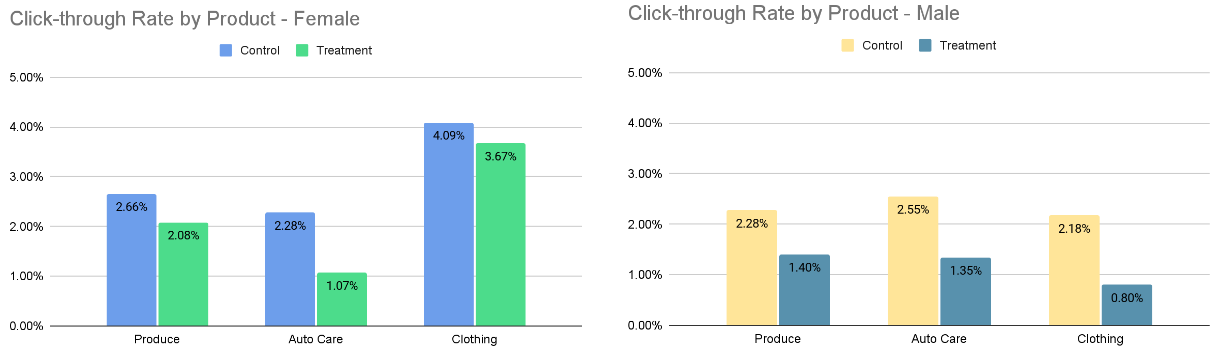


Figure 6: Click-through Rate Comparison by Product and Gender

Treatment Effect by Age

Although the magnitude of the treatment effect varied by age group, negative treatment effects were observed across the age groups. Audiences aged 65 years old and above saw the highest click-through rate in both control and treatment groups compared to the other age groups. The largest difference between treatment and control group was observed in the 35 to 44 year old group.

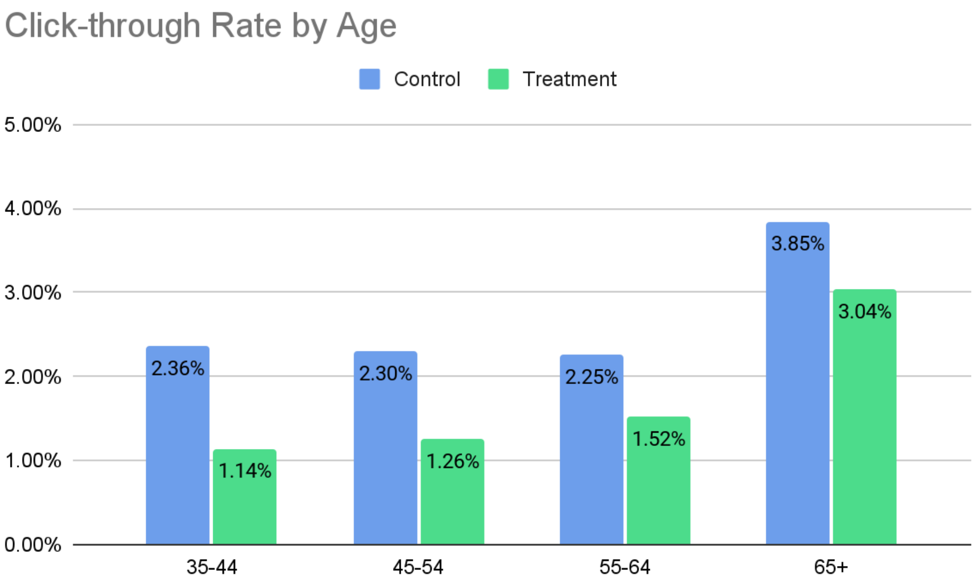


Figure 7: Click-through Rate by Age Group

While the magnitude of the treatment effect varies, negative treatment effects were observed consistently across product categories and age groups. For 35 to 44 year old audiences, seeing the language “All Natural” or “100% Natural” does not seem to impact the click through rate for produce. The minimal treatment effects were also seen in clothing for the audiences who are 45 years old or older. Auto care tended to see the largest negative treatment effect across the age groups.

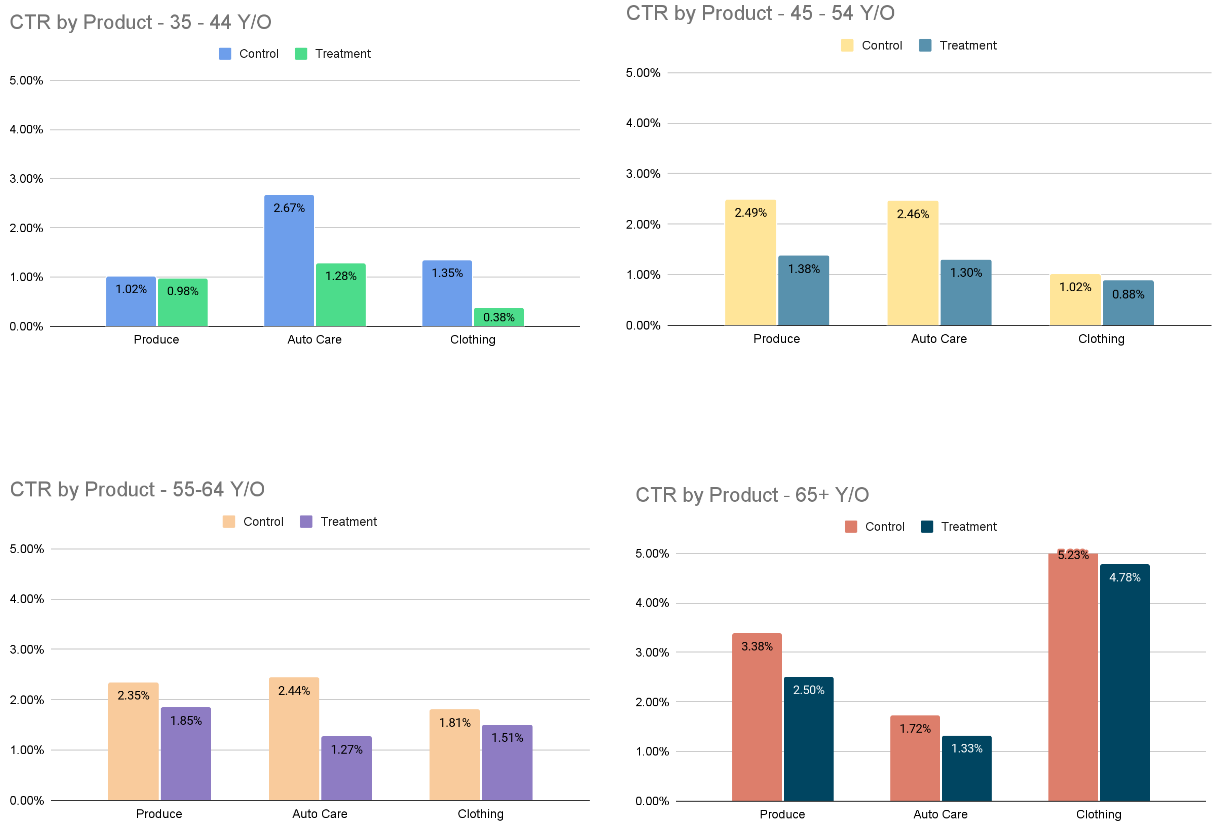


Figure 8: Click-through Rate Comparison by Product and Age Group

Impact of Randomization Issues

In order to understand how the covariate imbalance between treatment and control groups could have affected our results, we also broke the data down into the smallest subgroups possible to see if certain trends were over/underrepresented. It's possible that with an ineffective randomization protocol, certain subgroups with covariate combinations could have wildly different average treatment effects that would go unnoticed in the aggregate. Examining the average treatment effects of each subgroup would allow us to see if the slight negative effect of natural labelling in the aggregate also held within groups that had the same covariate values, or whether the imperfect randomization had masked more varied results.

Splitting the data into all possible combinations of covariates leaves us with 72 subgroups (36 treatment and 36 control) that yield 36 separate average treatment effects. Out of these 36 subgroups, only 5 yielded a positive ATE which contradicted the general trend of our results, and these 5 groups accounted for only ~11% of the total population involved in the experiment (10688 out of 93909). The 5 groups with different results also had a nearly identical ratio between treatment and control group sizes (0.80 to the other 31 subgroups' 0.79), suggesting they did not benefit from a more random randomization protocol.

This leaves our results in an unusual place. While Facebook's randomization protocol did not perform with the statistical rigor we had expected it to, it appears that the results we observe in the aggregate are largely mirrored by the results in the subgroups constructed from covariate combinations. Since subjects within these subgroups have identical covariate values, they are unaffected by the issue with the randomization protocol, so their mirroring of the findings in the aggregate suggest that our results may still represent a degree of truth despite the issues with treatment assignment. Further research on this problem would likely help us understand the validity of our results in this experiment.

Limitations

The most obvious limitations we faced in our experiment were not due to our design, but rather Facebook's algorithm for displaying advertisements to audiences. While we expected Facebook's A/B testing framework to produce control and treatment groups that could stand up to statistical scrutiny, we found the groups it produced contained imbalanced covariate distributions. Furthermore, we also found that certain age groups received little to no impressions on our advertisements, which impacted the completeness of our dataset. In order to combat these issues, we would have to either research an alternative advertising platform which allowed us better control of the randomization protocol and recruitment, or we would have to scale down our study to another format and run it entirely end-to-end by ourselves. However, a reformatted experimental design for end-to-end control would likely have to involve a questionnaire or interviews, which could bias the outcomes and move our results away from the true response of people to advertisements in everyday life.

If we were to keep the same general experimental design, there are still a few additional improvements we could make. The first possible improvement would be to improve our advertisements through the use of professional marketers or marketing consultants, as our ads may not have drawn in the full audience they could have. Using more effective advertisements could drive up click rates, which would reduce the effect of slight random noise on our results due to the larger numbers at play. In addition to improving the appeal of our advertisements, it could also be worthwhile to add a third category of advertisements that would mirror the layout of the treatment campaign, albeit with other additional text besides the “natural” label. Adding this third placebo campaign would allow us to parse out the effect of the word “natural” from the effect of simply adding more text/labels to advertisements.

Beyond the advertisements themselves, we could also expand the data available for our analysis by building a functional website that customers could explore. The absence of a functional website actually resulted in our advertisements being taken offline briefly during our data collection phase, but building out functionality could have more positive effects than just allowing for smoother data collection. Primarily, tracking clicks and additions to a shopping cart could give us even more fine-grain data on customer preferences relating to natural labelling beyond simply relying on advertisement clicks. Finally, the data in our experiment could be increased in scope even further by bringing in additional product categories for campaigns beyond the three we selected for this study. Examining the results of treatment on click through rates for more product categories could give us a much clearer sense of whether the treatment effect of natural labelling is largely uniform across different products or whether it varies heavily from industry to industry.

Conclusion

In our efforts to answer the research question of whether consumers have greater interest in products when they are labeled as being natural, we found several unexpected results. The first unexpected result centered around Facebook's A/B testing platform, which was not able to complete a randomization protocol that stood up to statistical scrutiny. We found evidence of significant covariate imbalances between treatment and control groups when conducting F-tests on our final data. The second unexpected result came when we began analyzing the average treatment effects of the natural labelling on the click through rate of customers. While we had intuited that natural labelling would increase the click through rate on our otherwise-identical advertisements, our data showed the exact opposite. This trend of the negative effect of natural labelling was maintained for each campaign, age group, and gender. Furthermore, an investigation into the average treatment effects for each subgroup of covariate combinations also yielded almost entirely negative treatment effects, suggesting that the randomization protocol may not have impacted our results as significantly as we initially thought.

These unexpected results leave us with additional questions for further research, and some ideas for how to address them. The first step in improving the validity of our results would be to implement a more rigorous randomization protocol. Facebook as a platform may not have the tools we need to randomize our treatment groups effectively, but we could explore other options as potential replacements. Ideally, we would be able to run this experiment end-to-end by ourselves, but doing so would likely require a scaling-down of the scope of the research, since our question hinges on large-scale consumer response to advertising campaigns. Other improvements would include professionally designed ads, placebo advertisements with other labels in place of “natural”, an increased number of product-specific campaigns, and a more robust storefront to gather additional data. We look forward to potentially implementing these changes in future research, but for now we'll look twice whenever an advertising campaign prominently displays the word “natural”.

References

[1] U.S. Food & Drug Administration. “Use of the Term Natural on Food Labeling.” *FDA Food Labeling & Nutrition*. Last updated Oct 22, 2018. Accessed Sep 14, 2021. https://www.fda.gov/food/food-labeling-nutrition/use-term-natural-food-labeling

[2] Hahnel UJ, Arnold O, Waschto M, et al. “The power of putting a label on it: green labels weigh heavier than contradicting product information for consumers' purchase decisions and post-purchase behavior.” *Front Psychol* (2015). 6:1392. Published 2015 Sep 23. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4585300/

[3] Kazlauskas, Benediktas. “Facebook CTR and How To Increase It.” *The Drum*. Published Jul 19, 2021. Accessed Oct 12, 2021. https://www.thedrum.com/profile/whatagraph/news/facebook-ctr-and-how-to-increase-it

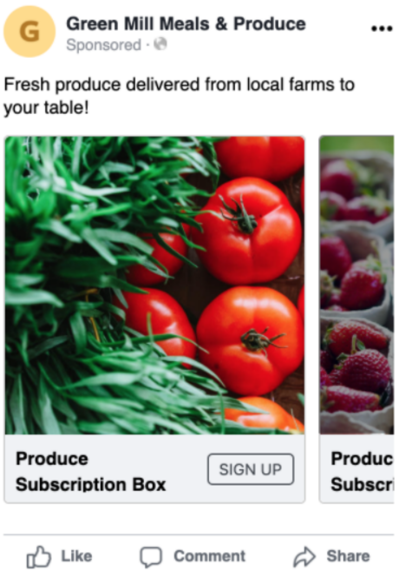
[4] Irvine, Mark. “Facebook Ad Benchmarks for Your Industry.” *WordStream*. Published Feb 20, 2020. Accessed Oct 12, 2021. https://www.wordstream.com/blog/ws/2019/11/12/facebook-ad-benchmarks

[5] ClinCalc. “Sample Size Calculator.” Accessed Dec 1, 2021. https://clincalc.com/stats/samplesize.aspx

Appendix A - Facebook Advertisements

Produce Ads

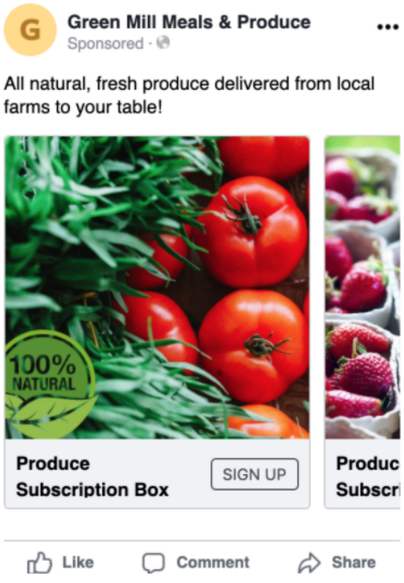
CONTROL



Green Mill Meals & Produce
Sponsored ·

Fresh produce delivered from local farms to your table!

TREATMENT

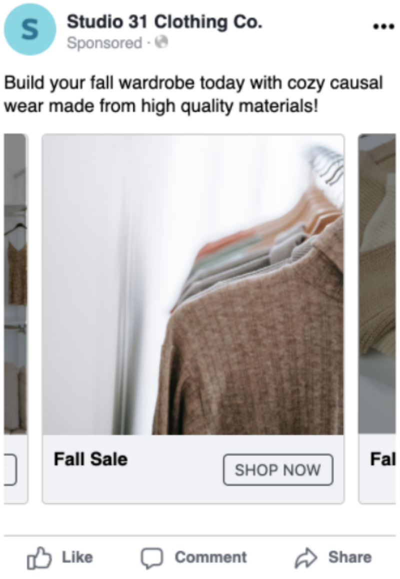


Green Mill Meals & Produce
Sponsored ·

All natural, fresh produce delivered from local farms to your table!

Clothing Ads

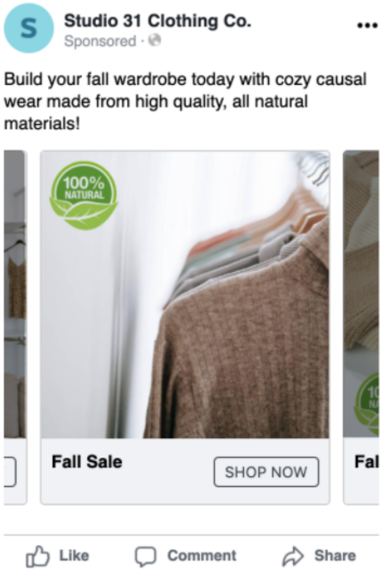
CONTROL



Studio 31 Clothing Co.
Sponsored ·

Build your fall wardrobe today with cozy causal wear made from high quality materials!

TREATMENT

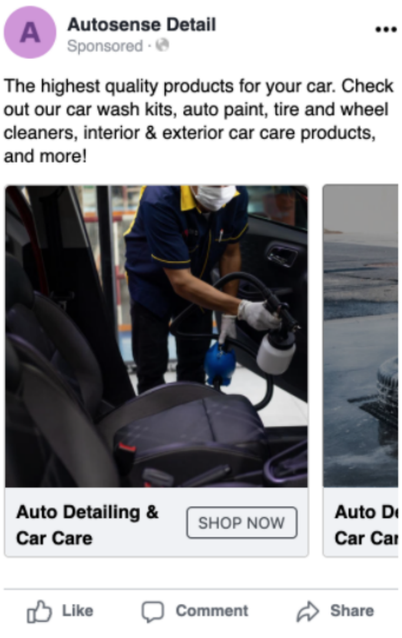


Studio 31 Clothing Co.
Sponsored ·

Build your fall wardrobe today with cozy causal wear made from high quality, all natural materials!

Auto Care Ads

CONTROL



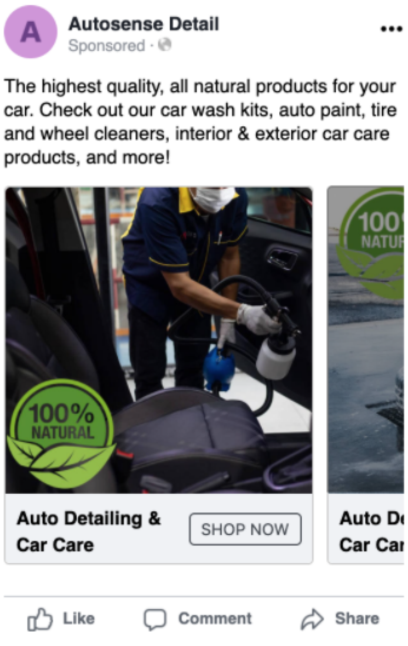
Autosense Detail

Sponsored

The highest quality products for your car. Check out our car wash kits, auto paint, tire and wheel cleaners, interior & exterior car care products, and more!

Auto Detailing & Car Care

TREATMENT



Autosense Detail

Sponsored

The highest quality, all natural products for your car. Check out our car wash kits, auto paint, tire and wheel cleaners, interior & exterior car care products, and more!

Auto Detailing & Car Care

Appendix B - Landing Pages for Fictitious Companies

Produce - Green Mill Foods
https://store-browsing.wixsite.com/greenmillfoods

Clothing - Studio 31 Clothing Co.
https://store-browsing.wixsite.com/studio31

Auto detailing - Autosense Detail
https://store-browsing.wixsite.com/autosense