THE UNIVERSITY OF CHICAGO

Leveraging Facebook for Behavioral Insights: The Influence of Algorithms on Experimental Research

By

Xinyi Zhang

06/01/2024

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Oleg Urminsky
Preceptor: Ryan Yuhao Fang

Abstract

This study investigates the influence of Facebook's algorithms on the outcomes of behavioral research experiments conducted on the platform. By analyzing a series of A/B tests, we highlight the existence of important but opaque underlying algorithms that perform targeting beyond demographic variables and how they affect the experimental results. The research primarily focuses on two types of campaign optimization options—click-optimized and view-optimized—and examines how these approaches influence the demographic composition of the audience and, consequently, the results of ad-effectiveness tests. Using a combination of logistic regression and chi-square tests, we provide empirical evidence that Facebook's algorithms significantly shape the substantial differences in both reach and click-through rates (CTR) between treatment and control groups, even after controlling for available demographics. The findings emphasize that experiments on Facebook do not operate under traditional random assignment, complicating the interpretation of treatment effects. This study contributes to our understanding of Facebook experimental results in social science research and the generalizability of its interpretations.

Keywords: Facebook experiments; A/B testing; algorithmic targeting; ad delivery; click-through-rate

1 Introduction

1.1 Background

In recent years, behavioral researchers have increasingly turned to online platforms like Facebook to conduct field experiments. The growing prevalence of algorithmically mediated environments for research makes them attractive for studying user interactions. However, this also introduces complexity. When running experiments involving search engines, social media platforms, or other algorithm-driven systems, external factors like content personalization and audience selection can influence results. This lack of control means that pure randomization isn't always guaranteed and experimental conditions can't be assumed to be equivalent on participant characteristics.

The key issue is that some behavioral studies make the mistake of treating Facebook and other algorithm-mediated platforms as controlled environments, like a laboratory setting. In the lab, a researcher can show a control group one version of an ad and a treatment group another, then directly measure their willingness to click. If the treatment ad leads to a 10% higher willingness to click, the increase in effectiveness can be ascribed to a difference in how people respond to the treatment, because all other factors are assumed to be roughly equivalent. However, in a Facebook experiment, launching a control and a treatment ad—even in an A/B test—and then observing a 10% higher click-through rate for the treatment ad does not necessarily mean people are more motivated by the treatment ad. Instead, the algorithm might have shown the treatment ad to individuals already predisposed to click more often, leading to a misattributed treatment effect. Thus, the conclusion that the treatment is better than the control (which would be valid in a lab test) might be flawed in an algorithmically mediated environment, because the experimental result is also influenced by the platform's differential targeting of the two ads, not just by the treatment itself.

This important yet often overlooked issue reveals a gap in understanding of whether (and how) algorithmic platforms like Facebook influence experimental results. While such experiments may be valid within-context (i.e., as long as the targeting algorithm does not change in relevant ways), the algorithmic targeting impedes the generalizability of results obtained in experiments conducted on Facebook to other contexts. Behavioral research and real-life practices need to recognize the complexities introduced by targeting algorithms in order to more critically think about experimental evidence and interpret its results.

1.2 Our Study

In our study, the overall objective is to test whether different ways of running campaigns on Facebook that involve different configurations of the targeting algorithm (specifically view-optimized vs. click-optimized) influence which participants receive which ad and, consequently, the experimental results. It should be noted that we are not attempting to identify specific elements (such as the headline or image in the advertisement) that contribute to the treatment effects. Instead, we compare the treatment ads against a corresponding control ad. Such experiments assess not only the effect of the treatment on Facebook user's preferences, but also the effect on the targeting decisions made by platform's algorithm for each ad. We test this by comparing different algorithms. For instance, when optimizing for click-through rates, if a treatment ad has a 10% higher click-through rate than the control, then our question is whether this 10% increase would also be observed when conducting the same experiment but instead optimizing for views. In other words, we are interested in whether changing the algorithm yields different insights from an experiment. To the degree that the experimental results differ when the algorithm differs, we will conclude that these experiments measure not only user responses but also the confounded impact of the algorithm.

Our hypothesis is that setting Facebook campaigns to optimize for clicks vs. to optimize for reach or views will yield different audiences due to differences in targeting. As a result, varying the optimization strategies (optimizing for clicks versus views) could subsequently impact the campaign's outcomes and overall conclusions from comparing a treatment to control group. Our analysis will consider available demographic factors of age, gender, and geographic location (state). Although these factors might not fully explain the differences in results, they can serve as indicators of whether the different algorithmic goals result in different targeting, influencing the conclusions drawn from the Facebook experiments run under that algorithmic goal.

We use **chi-square tests** of independence to specifically examine whether the distribution of ads across different demographic groups deviates more than expected by chance. By focusing on the metrics **'reach'**, our goal isn't to analyze the final results but to understand whether the algorithm is influencing which ads different demographic groups receive. Since we are interested in comparing Facebook experiments to a truly randomized experiment—where significant demographic differences across conditions should occur only about 5% of the time due to chance—this raises the question of whether Facebook experiments show more demographic differences than expected or align with this *5% false positive* rate. In some analyses, we'll also use a *Bonferroni correction* to adjust for multiple testing.

To understand if the algorithmic optimization goal and observed demographic differences translate to differences that significantly affect campaign results, we conduct **logistic regression analyses**. The primary outcome of interest is the treatment effect: the difference between the click-through rate for the treatment ad(s) compared to the control ad. The analyses will yield multiple treatment effects, each representing the difference in click-through rate between a treatment ad and the corresponding control ad in an A/B test. We conduct regression analyses with and without controlling for the available *demographic variables*. The objective would be to see whether controlling for these factors changes the results. We also conduct *likelihood ratio tests* as a test of whether treatment effects vary with the selection of optimization strategies overall, across the full set of ads.

To evaluate the role of targeting when running Facebook experiments and measure the degree to which it affects outcomes of the experiments, we first analyze demographic differences when different optimization goals are used to see if targeting plays a role. Then, we perform a regression analysis assessing the effect of ad differences on click-through rates that controls for demographic variables like age, gender, and region. If the measured treatment effects still differ more than expected by chance when controlling for demographics, targeting is influencing the results even beyond the available demographics. Thirdly, we compare the effects of campaigns optimized for clicks with those optimized for views. If the treatment effects significantly differ between the two algorithmic optimization strategies, we will conclude that the targeting algorithm does influence the campaign outcomes.

This thesis is organized as follows: In Section 2, we provide a review of the related literature. In Section 3, we introduce how our experiments were set up. In Section 4, we test whether there's a difference in targeting even when we use the A/B testing feature[1] on Facebook by several demographics comparisons where we always run separately for the click optimized and for the view optimized ads. Sub-sections 4.1, 4.2, and 4.3 test for demographic differences across ad executions (comparing between algorithmic optimization goals, both pooling all ads and within ad sets) and demographic differences across campaigns (comparing the identical-content control ads across different campaigns), and demographic differences between control-treatment ad pairs, respectively. In Section 5, we conduct multiple regression analyses to compare how treatment effects (click-through rate differences across ads) differ depending on the algorithmic optimization goal. Sub-section 5.1 contains logistic regressions on click-through rates performed separately within each ad set, both with and without demographic controls. Similarly, Sub-section 5.2 includes logistic regressions for pooled data across all ads, where we add 'ad set' level control, with and without controlling for demographic variables. Section 6 provides a summary of the study and discusses its contributions.

2 Prior Discussion

2.1 Historical Context of The Field

Discussion of the validity and reliability of online experiments (and the impact of targeting strategies) as a counterpart to traditional laboratory settings was initiated by a debate over an influential paper that used Facebook experiments.

Kosinski et al. (2015) demonstrated the potential of Facebook "Likes" to predict Big-Five personality attributes, providing a basis for targeted advertising. Extending this, Matz et al. (2017) conducted Facebook experiments on psychological profiling, showing that "ads tailored to the psychological profiles of users (inferred from their "Likes") could significantly enhance user engagement and purchasing behaviors. These studies integrated digital platform capabilities into behavioral research, suggesting the effectiveness of targeting strategies in online environments based on prior research on personality differences.

However, the internal validity of such online experiments was soon called into question. Eckles et al. (2018) raised concerns about the non-random assignment of online platforms, including Facebook, which could introduce significant confounds into experiments on algorithmic platforms. They highlighted that "Facebook's optimization algorithms might bias the user engagement metrics," thereby obscuring whether effects are genuinely due to the

[1] In this study, A/B testing will always refer to Facebook's A/B testing procedure, which is also known as Split Testing. It allows advertisers to test up to five different ads within a single experiment. Facebook randomly assigns users into different groups and shows each group only one variation of the ad, so that no user is exposed to multiple treatments in the experiment.

experimental manipulation or are artifacts of how the platform's underlying algorithms respond to the personality-based targeting. This critique pointed to a critical gap in the methodological rigor required for online behavioral studies, when the intent is to learn something generalizable about people's responses, not just platform-specific performance.

Matz et al. (2018) countered these criticisms by acknowledging the potential for algorithmic confounding but arguing that their experimental designs would "mitigate these effects sufficiently to validate our conclusions." Specifically, they incorporated controls for age, gender, and their interactions with the ad version into their analysis, confirming that the effects of psychological matching were robust even when these factors were accounted for. They highlighted that despite significant demographic differences among target groups, these differences had minimal impact on the results, and the effect sizes remained consistent when controls were included. This debate underscores the complexities of conducting rigorous behavioral research in online settings, and the difficulty in evaluating how sensitive experimental outcomes are to the external algorithmic influences.

2.2 Ongoing Discussions

The ongoing scholarly debate concerning the impact of algorithms on ad delivery and the outcomes of behavioral experiments on Facebook features diverse viewpoints but suffers from significant gaps. Only a handful of papers have addressed these issues, revealing a notable lack of comprehensive discussion and robust empirical evidence.

Kosinski et al. (2015) promote the use of Facebook for social science experiments, highlighting its ability to efficiently reach large and diverse populations. They acknowledge the platform's algorithmic mediation but focus on its potential for broad-scale data collection and the benefits of being able to target specific demographic groups. They argue that, despite potential biases, Facebook's robust data-gathering capabilities offer valuable opportunities for experimental research.

Ali et al. (2019) delve into the subtleties of Facebook's targeting algorithms through an innovative experiment. The researchers created variations in images using shades of white that were imperceptible to the human eye but recognizable by Facebook's algorithms. This manipulation affected which demographic groups saw the ads, indicating that Facebook's algorithms can detect and respond to such nuanced differences. The study also suggests that both view and click optimizations are subject to algorithmic biases that can skew ad delivery and impact the fairness of digital advertising practices.

Orazi and Johns (2020) advocate for optimizing Facebook experiments for views rather than clicks, under the premise that click optimization overly relies on Facebook's targeting algorithms, potentially skewing results. However, they base their recommendations on theoretical assumptions rather than empirical evidence, merely asserting without substantiation that view optimization circumvents algorithmic targeting. In fact, the results of Ali et al. (2019) suggest that it may not be possible to “turn off” Facebook’s targeting in a way that would allow true random assignment in an experiment.

Although the studies diverge in their conclusions, they uniformly highlight a significant challenge in using data from Facebook experiments for research: the opaque nature of its algorithms. Without a definitive understanding of the role these algorithms play, researchers cannot fully or accurately interpret their experimental results. This underscores the urgent need for further research in the field.

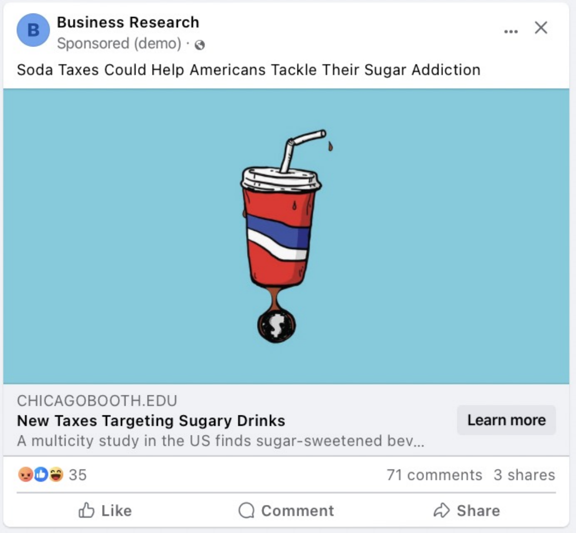
3 Set Up

Facebook introduced A/B testing, also known as split testing, in 2017 (Meta for Business, 2018, 2024). However, the A/B testing randomly assigns Facebook users prior to algorithmic targeting. As a result, Facebook’s A/B testing does not necessarily guarantee that who is served up the ads will be effectively random. In this research, we use the A/B testing feature to evaluate the degree to which the algorithmic goal affects measure treatment effects despite the use of the A/B testing feature.

In 2023, we conducted three A/B tests (we will call them A, B, C) with a total of ten treatment ads split across the three campaigns and the same control ad included in each campaign. We ran each A/B test two ways – either optimizing for clicks or optimizing for impressions (e.g., the number of views of the ads).

In 2024, we conducted two A/B tests (called D and E) with a total of five treatment ads split across the two campaigns and the same control ads in each campaign. We used a higher daily budget in 2024 and ran the campaigns and experiments for longer. We ran two versions of each campaign, either optimizing for clicks (as in 2023) or optimizing for reach (unique views), instead of total views as in 2023. In both years, we used a general Facebook audience population.

Our experimental setup involves testing ads for magazine articles, one article in 2023 and a different article in 2024. The articles were taken from the Chicago Booth Review, and the title and image on the Chicago Booth Review webpage for the article were adapted to be the control Facebook ad. MBA students, most of whom had prior marketing experience, worked in small groups to create their own alternative ad designed to perform better than the control ad, with each group of MBA students designing one ad, by modifying the text, image, and overall design. Because the treatment ads are named after students’ last names, we anonymize these names in the tables in this study. We refer to the ads as A1, A2, ..., B1, B2, etc. Figure 1 is the preview of the control ad in 2024, and Figure 2 is a preview of one of the ten treatment ads designed by one group of MBA students. The setup simulates



Business Research

Sponsored (demo) ·

Soda Taxes Could Help Americans Tackle Their Sugar Addiction

CHICAGOBOOTH.EDU

New Taxes Targeting Sugary Drinks

A multicity study in the US finds sugar-sweetened bev...

Figure 1: Preview: Control Ad (2024)

how firms often develop and test multiple versions of ads internally before launching a full campaign.

Facebook only provides summary data on the outcome of each campaign. A significant limitation of current reporting from Facebook is that the two available reports provide demographic breakdowns by state and by age-gender groups only, i.e., not individual-level data and not based on any other demographic or psychographic categories. The age-gender groups are predefined by Facebook and cannot be modified.

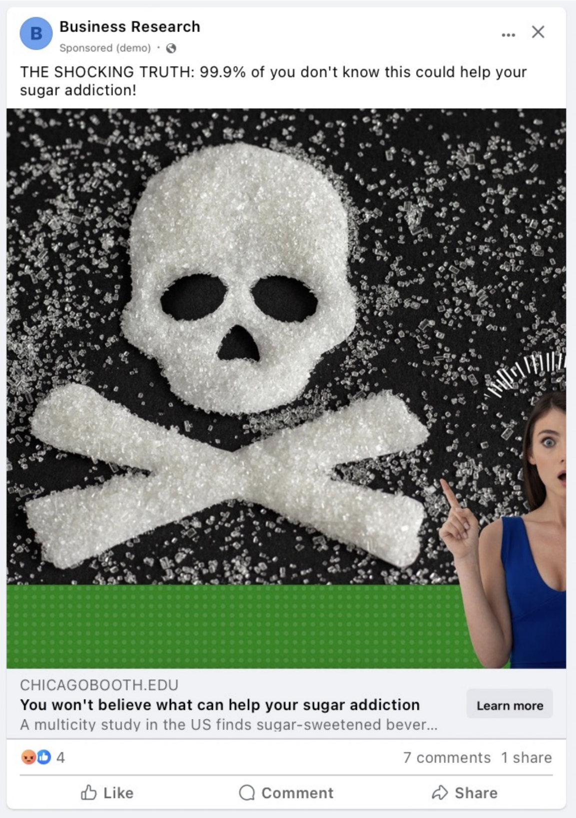
4 Does Targeting Exist for View/Reach- and Click-Optimization?

4.1 Demographic Test: Optimization Strategy

We explore whether the demographic composition of ad audiences differs when optimizing for clicks versus views. We analyze the distribution of views of advertisements across different age, gender, and region demographic groups.

4.1.1 Part 1: Aggregated Analysis

Initially, we conduct Chi-square tests to assess whether, in general, advertisements across ad sets with a click-optimization goal selectively target demographic groups differently compared to those with a view-optimization goal. In other words, we aim to determine if the two optimization strategies result in ads reaching different proportions of demographic audiences. We sum up the reach counts for all ads within each demographic group, separately



Business Research

Sponsored (demo) ·

THE SHOCKING TRUTH: 99.9% of you don't know this could help your sugar addiction!

CHICAGOBOOTH.EDU

You won't believe what can help your sugar addiction

A multicity study in the US finds sugar-sweetened bever...

Figure 2: Preview: Example Treatment Ad (2024)

for each optimization goal. For example, in Table 3, the entry in the row '18-24 female' and column 'ViewOpt' displays the sum of reach for all campaigns that optimize views and are distributed to females aged 18-24 in 2023. We have 4 such tables and conducted corresponding Chi-square tests. The p-values are shown in Table 1. These results show that when optimizing for clicks, ads reach a different demographic composition compared to when optimizing for views. This distribution would be extremely unlikely to occur by chance, based on the statistical significances.

Test (Column x Row)	**Chi-Square Statistic**	**p-value**	**DoF**	**Data Year**
Optimization x Age-Gender	10580.81	0.0	11	2023
Optimization x Age-Gender	32191.94	0.0	11	2024
Optimization x Region	1016.34	$< .01$	50	2023
Optimization x Region	2507.73	0.0	50	2024

Table 1: Chi-square Test Results: Reach Across Optimization Strategy (All Ads Combined)

Our results show that demographic differences in reach exist between view-optimized and click-optimized campaigns. This discrepancy could indicate that one strategy is using targeting while the other is not, or it could mean that both strategies employ different targeting, either of which results in demographic differences.

In other words, different versions of the targeting algorithm result in significant differences across campaigns. Thus, comparisons between campaigns are influenced by the specific algorithm version used.

4.1.2 Part 2: Within Ad Sets

The previous analyses provide overall tests of the demographic discrepancies between view-optimized and click-optimized campaigns. To determine whether both strategies employ targeting algorithms, we analyze whether demographic composition differs within each ad set.

Methodology We conduct separate chi-square tests for different demographic categories across different optimization strategies, as follows:

(1) Chi-Square Tests by Age-Gender Categories We compared the reach across ads (control and treatment ads) within a set by age-gender categories. This allows us to explore whether demographic differences arise between ads based on their targeting strategies (click optimization vs. view optimization) across age-gender groups.

(2) Chi-Square Tests by Region Categories We also conduct chi-square tests comparing the reach across ads (control and treatments) within a set by region categories. This helps to

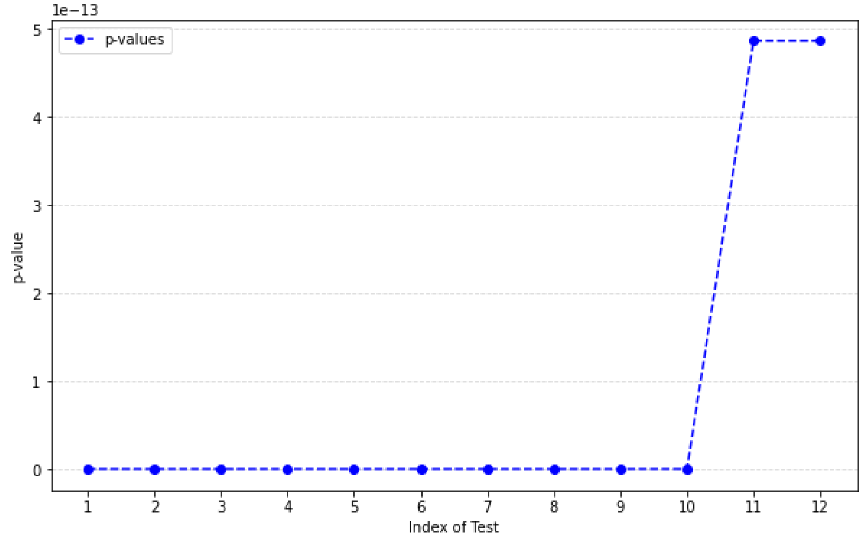


Figure 3: P-values of Chi-Square Tests (Sets in 2023)

determine if demographic differences emerge between ads based on their targeting strategies across different regions.

Summary of Tests Table 4 is an example of the contingency tables analyzed in this part, showing the reach by age/gender categories across one control and four treatment ads that were all included in a single experiment. In 2023, a total of 12 chi-square tests were conducted across three sets (A, B, and C), examining both click- and view-optimized, and covering both age-gender and region categories. In 2024, 8 chi-square tests were conducted across two sets (D, E), similarly examining clicks and views, and covering both age-gender and region categories.

Results and Interpretation The 20 chi-square tests all yielded significant results, with 20 near-zero p-values. The distribution of p-values are presented in Figures 3 and 4. This indicates that there are demographic differences across the ads within each experiment.

The significant demographic differences across ads within an experiment are observed for all experiments, including those that optimized for clicks as well as those that optimized for either reach or views. This is clear evidence that demographic targeting exists in both algorithmic optimization strategies.

Thus, we provide evidence that refutes the claim by Orazi and Johns (2020) that view optimization does not use an algorithm in Facebook, and that instead supports the understanding that Facebook uses a targeting algorithm for both click-optimized and view-optimized (or reach-optimized) strategies. There's no way to address the issue of audience

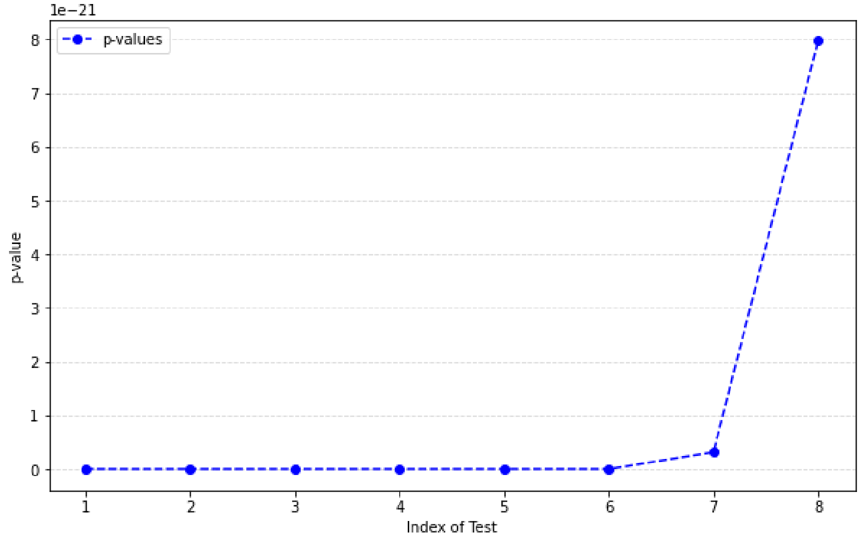


Figure 4: P-values of Chi-Square Tests (Sets in 2024)

variation simply by selecting a specific optimization algorithm, as either will result in different demographic distributions across the ads within an experiment.

4.2 Demographic Test: Campaign Contents

In this analysis, we aim to determine whether the observed demographic differences in reach (as shown in Section 4.1.2), can be explained by the content of the ads themselves. If one optimization algorithm strictly targets based on the content of the ad itself, then identical ads in different experiments (with the same algorithmic optimization goal) should have roughly the same reach across demographic groups.

Methodology We conducted Chi-square tests to compare the demographic composition of the audience across the identical-content control ads included in different experiments, all of which were optimizing for the same goals. This analysis focuses specifically on the demographic targeting impact of the ad content itself.

Summary of Tests Table 5 is an example of the contingency tables analyzed in this part. For data in 2023, we performed a total of 4 Chi-square tests, covering both Age-Gender and Region categories, for both click- and view-optimized strategies. For data in 2024, we repeated this process, again performing 4 Chi-square tests.

Results and Interpretation The results are presented in Table 2. Five of the 8 tests were not significant at 5% level. In 2023, all 4 tests were not significant, indicating that the

Test (Column x Row)	**Chi-Square Statistic**	**p-value**	**DoF**	**Data Year**	**Optimization**
Control Ads x Age-Gender	25.89	0.26	22	2023	OptView
Control Ads x Age-Gender	19.74	0.59	22	2023	OptClick
Control Ads x Region	117.31	0.11	100	2023	OptView
Control Ads x Region	120.01	0.08	100	2023	OptClick
Control Ads x Age-Gender	520.91	$< .01$	11	2024	OptReach
Control Ads x Age-Gender	10.41	0.49	11	2024	OptClick
Control Ads x Region	822.27	$< .01$	50	2024	OptReach
Control Ads x Region	455.57	$< .01$	50	2024	OptClick

Table 2: Chi-square Test Results: Reach Across Identical-Content Control Ads

control ads within each set have no significantly different audience demographic composition when they share the same optimization strategy (views or clicks). In 2024, however, only 1 out of 4 tests was not significant. Specifically, when optimizing for clicks, the two control ads had no significantly different audience in terms of reach across age/gender.

As explained earlier, if the algorithms strictly targeted based on the content of the ad itself, then identical ads in different sets should reach roughly the same demographics. However, 2024 data shows that even with identical content, demographic differences often still occurred.

One possibility is that Facebook's algorithms may target based not only on the ad's content but possibly on the entire ad set as well. This would explain why, even when the same control ad was used across different ad sets, we observed differences in the reach across demographic groups. Another explanation could be that the algorithm is sensitive to early interactions. For instance, if more older users initially click on the control ad in one experiment compared to the same control ad in another experiment, the algorithm might target more older users in the first experiment, but would target younger users in the second experiment, leading to differences in reach by demographics for the same ad in different experiments.

The key takeaway is that advertising on Facebook is not like a controlled experiment where people are randomly assigned to ads. Different versions of the algorithm, variations in ad content, and even the experiment an ad is included in can cause significant demographic differences. This discrepancy across the same control ad suggests that factors beyond the ad content itself, such as ad set or algorithmic sensitivity to early interactions, may play a role.

4.3 Direct Demographic Test: Control-Treatment Pairs

This part is aimed to provide the most direct examination of our hypothesis. We would like to understand if the algorithm treats treatment and control ads differently, by checking if it leads to different audience compositions for each.

Methodology *(1) Chi-Square Tests by Age-Gender Categories* We compare the reach across age-gender groups of each treatment ad to the reach across age-gender groups for the control ad within the same ad set (experiment).

(2) Chi-Square Tests by Region Categories We compare the reach across U.S. states for each treatment ad to the reach across U.S. states for the control ad in the same ad set (experiment).

(3) Bonferroni Correction Because we are conducting multiple tests, the overall false-positive rate can be inflated. We adjust for the family-wise error rate (FWER) by applying a Bonferroni correction afterward, setting the significance level at 0.05 divided by the number of tests, for each year's tests.

Summary of Tests Table 6 is an example of one of the contingency tables analyzed in this part. For data in 2023, 40 such tests in total were conducted, spanning 10 treatment-control pairs in ad sets (4 in A, 4 in B, 3 in C). These tests examined both clicks and views, covering age-gender and region categories. The Bonferroni threshold was 0.00125. Similarly, for data in 2024, a total of 20 tests were conducted, spanning 5 treatment-control pairs in ad sets (3 in D, 2 in E). These tests also examined both clicks and reach, covering age-gender and region categories. The Bonferroni threshold was 0.00250. When combining data from 2023 and 2024, the total number of tests conducted was 60, spanning 15 pairs in ad sets (4 in A, 4 in B, 3 in C, 3 in D, 2 in E), also examining clicks and exposure[2], covering age-gender and region categories. The Bonferroni threshold was 0.00083.

Results and Interpretation In 2023, 34 of the 40 tests showed significant results post-Bonferroni correction, as shown in Figure 5. In 2024, all tests were significant after correction, as shown in Figure 6. Cumulatively, as shown in Figure 7, 54 out of 60 tests over the two years were significant, i.e. overall 90% are significant after Bonferroni correction.

The analysis tests whether there are demographic differences between each treatment ad and the control ad in the same experiment, and about 90% of the time, there are notable disparities. This reinforces our overarching point that one cannot compare treatment and control groups and treat the process like a traditional experiment because the groups consist of systematically different people, and the assignment to different ads within an experiment is not random, because the algorithm is treating the groups differently.

Another key question is whether more demographic differences are coming from the clicks-optimization or from the views/reach-optimization. The distribution of the 6 insignif-

[2] As 'exposure' encompasses both the concepts of 'view' and 'reach', we employ it to collectively refer to them in our combined analysis for the years 2023 and 2024.

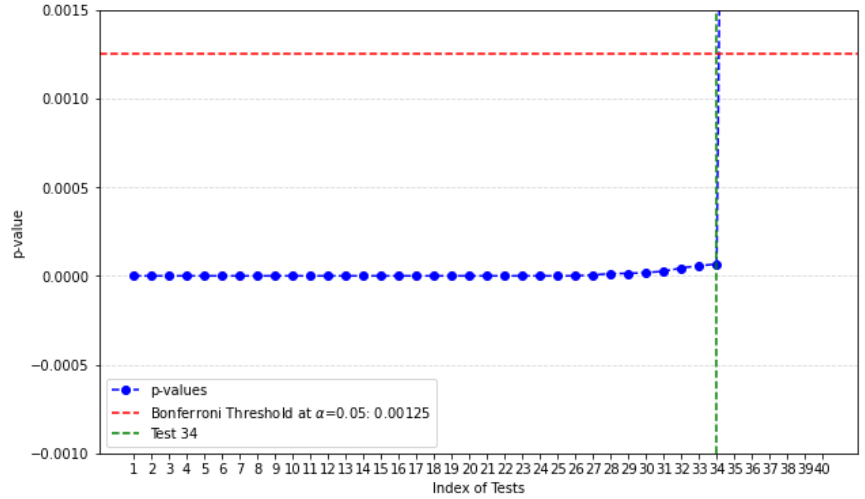


Figure 5: P-values and Bonferroni Threshold (Sets in 2023)

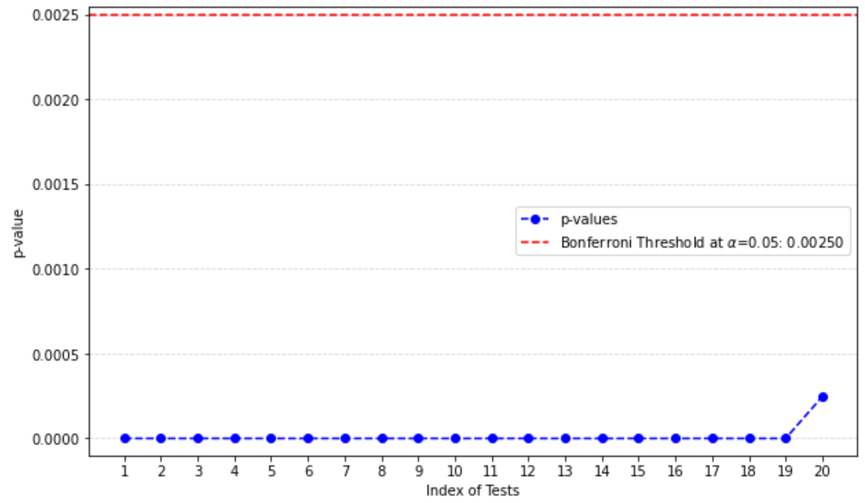


Figure 6: P-values and Bonferroni Threshold (Sets in 2024)

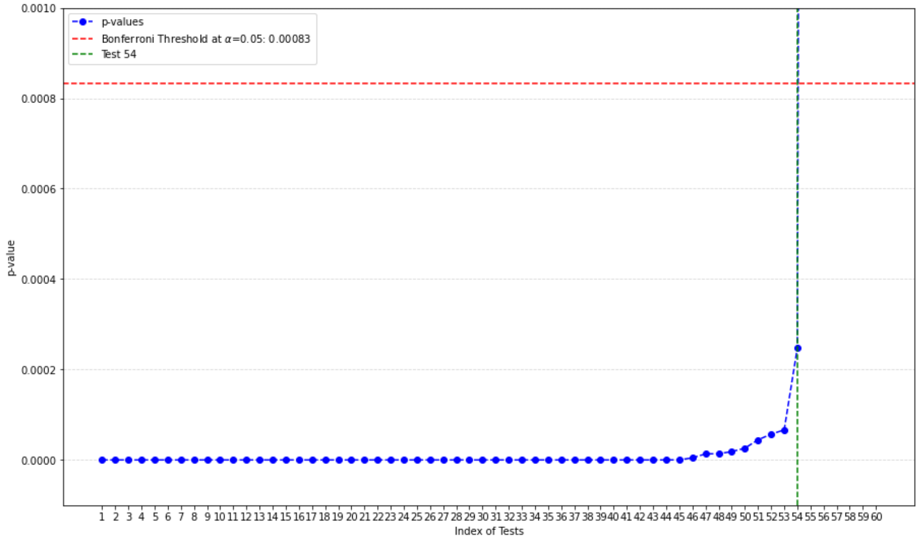


Figure 7: P-values and Bonferroni Threshold (Sets in 2023 and 2024)

icant results (two each across age-gender views, region clicks, and region views) suggests a relatively even distribution of failures across both click- and view/reach-optimized strategies. This even distribution implies that there is no significant difference between strategies optimized for clicks versus views in terms of demographic balance between treatment and control ads. These findings suggest that both optimizing for clicks and for views/reach would involve targeting algorithms that would show different ads to different people. This systematic targeting affects the assignment process and complicates the interpretation of treatment effects for both algorithmic optimization strategies.

5 Does Targeting Matter? - CTR comparison across ads

Analyses in Section 4 establish that both views (or reach) and clicks involve algorithmic targeting. Next, we examine the effect of the observable algorithmic targeting (i.e., in terms of age/gender and region) on ad performance. The key question addressed is whether the differences in demographic targeting between optimizing for clicks vs. for views/reach translate into different treatment effects.

5.1 Within Each Ad Set

5.1.1 Part 1: Baseline Model

In this part, we analyze click-through rates (CTR) across various advertisements, broken down by sets of ads (i.e., distinct experiments) labeled A, B, C, D, and E. Essentially, we run separate regressions, one for each set. Our aim is to examine whether treatment effects (measured by comparing CTR for treatment vs. control ads) differ depending on the algorithmic optimization goal.

Methodology:

From the summary data provided by Facebook, we construct person-level datasets that match on summary statistics. A logistic regression model is then fit on this data to predict whether a unique click occured (1) or not (0). The model incorporates three key variables as predictors:

1. Ad type: This distinguishes between control and treatment ads.
2. Campaign type: Whether the campaign is optimized for clicks or views/reach.
3. Interaction terms: The interaction between ad type and campaign type to explore if the treatment effect on CTR differs based on campaign type.

The model equation can be represented as follows:

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{N} \beta_i \cdot \text{Ad}_i + \beta_C \cdot \text{Campaign}_C + \sum_{i=1}^{N} \beta_{iC} \cdot (\text{Ad}_i \times \text{Campaign}_C) \qquad (1)$$

Where:

- $\text{logit}(p)$: The log-odds of a unique click occurring.
- Ad_i: A categorical variable representing the ith treatment ad, with the control ad as a reference.
- Campaign_C: A categorical variable representing campaign type, with ”Optimized for clicks” as a reference.
- $\text{Ad}_i \times \text{Campaign}_C$: An interaction term between ad type and campaign type.
- β_0: The intercept term.
- β_i: The main effect for the ith treatment ad, representing the difference in click-through rate between the ith treatment ad and the corresponding control ad.
- β_C: The main effect for campaign type, representing the difference in click-through rate between ads optimized for views and clicks.
- β_{iC}: The coefficient for the interaction between the ith treatment ad and campaign type, representing the difference in click-through rate for the specific ith treatment ad when optimizing for views compared to clicks.

First, we use logistic regression to check for a difference in CTR between control and treatment ads. Then, we include campaign type to account for potential CTR differences between click-optimized and view-optimized campaigns. This inclusion allows us to control for the level differences between campaigns optimized for clicks versus those optimized for views. As the last step, we incorporate an interaction term between ad type and campaign type, enabling us to explore whether the treatment effect on CTR varies based on campaign optimization strategy.

After fitting the models, likelihood ratio tests are conducted comparing the full model with the interaction term to a simplified model without interaction terms, as an overall test of whether treatment effects differ by campaign type.

Results: The regression results are shown in Tables 7, 8, 9, 10, 11.

(1) CTR Across Campaign Types: Ads optimized for views or reach versus those optimized for clicks yield substantially lower click-through rates.

(2) Treatment Effects and Interaction Terms:

- Set A (2023): One treatment had a negative effect, two had positive effects, and one had no significant effect. One out of four interactions was significantly positive.
- Set B (2023): One treatment had a negative effect, one had a positive effect, and one had no significant effect. One out of three interactions was significantly negative.
- Set C (2023): Two treatments had positive effects, and one had no significant effect. One out of three interactions was significantly positive.
- Set D (2024): Three treatments had positive effects. None of the three interactions was significant.
- Set E (2024): One treatment had a negative effect, one had a positive effect. None of the two interactions was significant.

(3) Likelihood-Ratio Tests: Table 12 displays the results. Three of the five ad sets showed statistically significant differences. Overall, four out of the 15 interaction terms returned significant results, combining data across both years. Three of the five Likelihood-Ratio Tests were (highly) significant. These results suggest that while treatment effects do not always significantly vary with the campaign type, they do so often enough that adding interaction terms to the baseline models significantly improves fit. In sum, differences in treatment effects are observed more often than would be expected solely by chance.

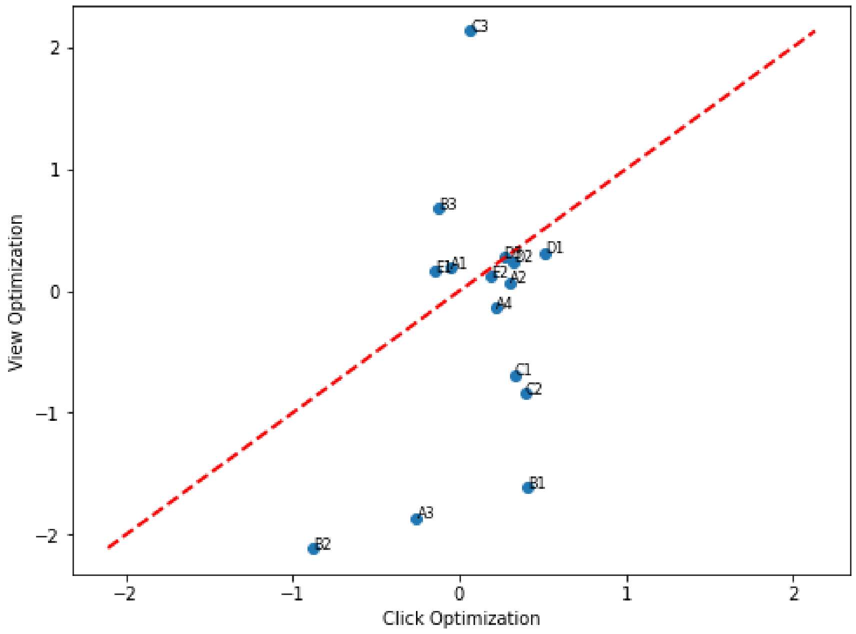


Figure 8: Scatter Plot: Treatment Effects Across Optimization Goals (No Control)

Interpretation Overall, experiments that aim to optimize for clicks can produce different results compared to those aiming to optimize for views. The treatment effects, as reflected in the coefficients, can vary significantly. Ads optimized for views versus those optimized for clicks yield lower click outcomes, highlighting the fact that algorithmic targeting does impact click-through rates. Four out of 15 interactions were significant, indicating that, without controlling for demographics, we observe significantly different results depending on whether experiments are optimized for clicks or views/reach.

To illustrate the observed differences in treatment effects, we *estimate the treatment effects separately, optimizing for clicks and optimizing for views/reach*, and plot the coefficients, both as a scatter plot and as paired histograms. In the scatter plot, if the coefficients for clicks and views align linearly, this suggests that whichever treatment performed better when optimized for clicks also performed better when optimized for views/reach. Alternatively, a lack of correlation could indicate discrepancies, such as one treatment being effective when optimizing for clicks but not when optimizing for views/reach (or vice versa). Through the paired histograms, we want to see whether a particular treatment might yield different directional conclusions (positive vs. negative effects) when comparing clicks versus views/reach. It could also help identify whether optimizing for clicks typically produces larger or smaller effects compared to optimizing for views/reach.

The scatter plot in Figure 8 shows a weak relationship between the treatment effect sizes of click optimization and view/reach optimization. A given effect size measured through

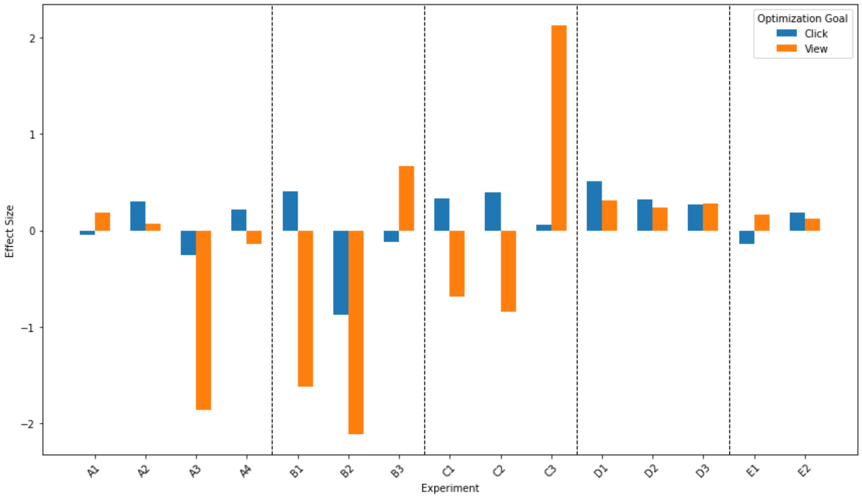


Figure 9: Paired Histogram: Treatment Effects Across Optimization Goals (No Control)

click optimization can result in a wide range of effect sizes measured through view/reach optimization, and vice versa. In some cases, the effect sizes even move in opposite directions. Indeed, the plot reveals a systematic difference, with more points below the line than above it, indicating weaker or more negative effects when optimizing for clicks compared to when optimizing for views/reach.

The paired histogram in Figure 9 is also quite revealing. It's intriguing that, while there are noticeable differences among the sets, the three ads in Set D appear to be quite similar to one another, possibly just by chance. However, we don't have (and don't need) an explanation for these differences. The main takeaway is that there is substantial variation, which means that optimizing for clicks versus optimizing for views can yield entirely different results. Specifically, a treatment ad could yield effects in opposite directions when optimizing for views versus clicks, and optimizing for views/reach often results in more extreme treatment effects, while click optimization tends to produce more modest effects in either direction. This reinforces the importance of carefully interpreting experimental outcomes, as the optimization strategy itself can significantly influence the results.

It is also noteworthy that the differences based on algorithmic optimization goal were smaller in 2024, although it is not clear whether that is because of platform changes, the larger sample size or the fact that reach was optimizing in place of views in 2024.

Overall, it is clear that optimizing for clicks or views/reach can yield different estimates of treatment effects. The next step is to test whether we can identify factors driving these differences.

5.1.2 Part 2: Controlling for Demographics

When optimizing for clicks, we find a quite different treatment effect than when optimizing for views. This may occur due to systematically different demographics across treatment and control ads, that vary between optimizing for clicks vs. for views/reach. Thus, the core question we address here is whether the significant interactions observed in the Equation 1 of the study, which pertain to campaign optimization, persist after we account for demographic differences.

Comparing Regression 1 to the ones that include demographic variables, if the interaction effect diminishes or disappears when these controls are added, it indicates that the optimization strategy affects treatment effects due to the specific demographic differences we observe (e.g., age-gender and U.S. state). If a significant interaction remains, it suggests that optimization type influences results beyond age, gender, and region.

Methodolody: *Logistic Regression:* The previous regression 1 in Section 5.1.1 includes the main effect of ads, the main effect of campaign type, and their interaction. In this part, our regression also includes these variables, and adds either age and gender or U.S. state as additional controls. Because Facebook does not provide a joint distribution of age, gender and U.S. state, we cannot include both age-gender and state within the same regressions.

The models can be represented as follows:

$$\begin{aligned}\text{logit}(p) = \beta_0 &+ \sum_{i=1}^{N} \beta_{\text{Ad},i} \cdot \text{Ad}_i + \beta_{\text{C}} \cdot \text{Campaign}_C + \sum_{i=1}^{N} \beta_{\text{Ad}C,i} \cdot (\text{Ad}_i \times \text{Campaign}_C) \\ &+ \sum_{j=2}^{6} \beta_{\text{Age},j} \cdot \text{Age}_j + \beta_{\text{Gender,M}} \cdot \text{Gender}_M\end{aligned} \tag{2}$$

$$\begin{aligned}\text{logit}(p) = \beta_0 &+ \sum_{i=1}^{N} \beta_{\text{Ad},i} \cdot \text{Ad}_i + \beta_{\text{C}} \cdot \text{Campaign}_C + \sum_{i=1}^{N} \beta_{\text{Ad}C,i} \cdot (\text{Ad}_i \times \text{Campaign}_C) \\ &+ \sum_{k=2}^{51} \beta_{\text{State},k} \cdot \text{State}_k\end{aligned} \tag{3}$$

Where:

- Age_j: A categorical variable representing age group j, where groups are 18-24, 25-34, 35-44, 45-54, 55-64, and 65+, with 18-24 as the reference.
- Gender_M: A categorical variable representing gender (Male), with Female as the reference.

- State_k: A categorical variable representing each state k, covering 51 U.S. states, including Washington D.C.
- $\beta_{\text{Age},j}$: The coefficient for each age group j compared to being age 18-24.
- $\beta_{\text{Gender,M}}$: The effect of being a male compared to being a female.
- $\beta_{\text{State},k}$: The coefficient for state k.

Note: the explanation of other terms in the equation is exactly the same as that in Equation 1.

Results The regression results can be found in Tables 13, 14, 15, 16, 17.

(1) CTR Across Campaign Types: Ads optimized for views versus those optimized for clicks yield lower click outcomes.

(2) Treatment Effects and Interaction Terms:

- Set A (2023): One treatment had a negative effect, two had positive effects, and one had no significant effect. One out of four interactions was significantly positive.
- Set B (2023): One treatment had a negative effect, one had a positive effect, and one had no significant effect. One out of three interactions was significantly negative.
- Set C (2023): Two treatments had positive effects, and one had no significant effect. One out of three interactions was significantly positive.
- Set D (2024): Three treatments had positive effects. None out of three interactions was significantly positive or negative.
- Set E (2024): One treatment had a negative effect, and one had a positive effect. None of the two interactions was significantly positive or negative.

Interpretation Both with and without demographic controls, the study's findings remain consistent, based on the likelihood ratio tests and tests of significant interaction terms. Significance was observed in 4 out of 15 interaction terms even after controlling for demographics. Likewise, for three of the five sets, the likelihood ratio tests were significant even after controlling for demographics.

This provides pretty strong evidence that the choice of optimization strategy in advertising campaigns—whether for views/reach or for clicks—significantly affects the outcomes. And while Facebook's disclosed demographic differences (age, gender and state) may be useful as indicators that the algorithms treat different ads differently, the results cannot be explained solely by targeting based on these demographic factors alone. Instead, the algorithm is likely incorporating numerous other characteristics that lead to the observed differences in treatment effects.

5.2 Pooling across the Ad Sets

The goal of this final section is to consolidate across all the ad experiments, to help draw general conclusions. By pooling all the data into one comprehensive analysis, we aim to observe if notable differences exist in treatment effects depending on whether campaigns are optimized for exposure or clicks.

5.2.1 Part 3: Baseline Model

In this part, we first run one overall regression for each year instead of separate regressions for each ad set. Ad sets A, B, and C from 2023 are analyzed together, as are ad sets D and E from 2024. Then, we combine the data from 2023 and 2024 all into one analysis, as we don't see major differences between these two years.

A fixed effect for each experiment is included in the regression, capturing the difference in CTR across the multiple controls. These differences are not of primary interest, but allow us to compare each treatment to all three controls. Taking the ad sets in 2023 as an example, the intercept represents the CTR for set A, the set B fixed effect indicates the difference between set B and set A, and the set C fixed effect indicates the difference between set C and set A.

Methodology *(1) Logistic Regression:* We use a logistic regression model to predict the probability of a unique click on an advertisement based on several key factors. Except for those considered in Equation 1, this equation additionally considers the ad set as a categorical variable, accounting for each ad set's distinct effect on the click probability. Essentially, we are pooling across ad sets to analyze the interactions between treatments and campaign type, as a measure of whether treatment effects differ by campaign type.

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{N} \beta_{\text{Ad},i} \cdot \text{Ad}_i + \beta_{\text{C}} \cdot \text{Campaign}_C + \sum_{i=1}^{N} \beta_{\text{Ad}C,i} \cdot (\text{Ad}_i \times \text{Campaign}_C) + \sum_{l=1}^{M} \beta_{\text{Set},l} \cdot \text{Set}_l \tag{4}$$

Here,

- Set_l: Set_l serves as a categorical variable representing the lth ad set. For individual year analyses, the reference ad set changes: in 2023, set A is used as the reference, whereas in 2024, set D serves this role. When conducting regression analysis that combines data from both 2023 and 2024, set A is the reference.

- $\beta_{\text{Set},l}$: The coefficient for the lth ad set, representing its fixed effect.

Note: (1) all control ads are coded as the same ‘control ad’ in this analysis. (2) the explanation of other terms in the equation is exactly the same as that in Equation 1.

(2) Likelihood Ratio Test: This test compares the full model (including interaction terms) with a simplified model that excludes these interactions to assess their significance.

Results Tables 18 and 19 present the regression results of running 2023 data and 2024 separately, and Table 20 is the result of the combined big regression. Table 21 presents the results of the likelihood ratio test.

(1) CTR Across Campaign Types: Ads optimized for views versus those optimized for clicks yield lower click outcomes.

(2) Treatment Effects and Interaction Terms:

2023 Analysis: Out of 10 treatments, two had negative effects, five had positive effects, and three showed no significant effect. Regarding interaction terms, five were significantly negative, and one was significantly positive.

2024 Analysis: Out of 5 treatments, two had negative effects, and three had positive effects. One interaction term was significantly positive.

2023 and 2024 Combined Analysis: Out of 15 treatments, three had negative effects, and nine had positive effects. Five interaction terms were significantly positive, two were significantly negative.

(3) Likelihood-Ratio Test: Both analyses showed significant results with p-values indicating strong statistical significance (2023: p=0.0000; 2024: p=0.0029).

Among the 15 interaction terms analyzed, 46.7% are statistically significant, including 6 out of 10 from 2023 and 1 out of 5 from 2024. This result holds when we conduct the comprehensive combined analysis for 2023 and 2024. The significant likelihood ratio tests confirm that overall, treatment effects differ by the algorithmic optimization goal more than would be expected to occur by chance.

Interpretation One key takeaway from this combined analysis is that overall, treatment effects differ based on the algorithm, but not consistently in one direction, making it difficult to predict the direction of the variation.

5.2.2 Part 4: Controlling for Demographics

In this part, we conduct a similar analysis as in Part 3, but with additional demographic controls for age, gender, and region.

Methodology *(1) Logistic regression*

The regression equations are as follows:

$$\begin{aligned}\text{logit}(p) = \beta_0 &+ \sum_{i=1}^{N} \beta_{\text{Ad},i} \cdot \text{Ad}_i + \beta_{\text{C}} \cdot \text{Campaign}_C + \sum_{i=1}^{N} \beta_{\text{Ad}C,i} \cdot (\text{Ad}_i \times \text{Campaign}_C) \\ &+ \sum_{j=2}^{6} \beta_{\text{Age},j} \cdot \text{Age}_j + \beta_{\text{Gender,M}} \cdot \text{Gender}_M + \sum_{l=1}^{M} \beta_{\text{Set},l} \cdot \text{Set}_l\end{aligned} \tag{5}$$

$$\begin{aligned}\text{logit}(p) = \beta_0 &+ \sum_{i=1}^{N} \beta_{\text{Ad},i} \cdot \text{Ad}_i + \beta_{\text{C}} \cdot \text{Campaign}_C + \sum_{i=1}^{N} \beta_{\text{Ad}C,i} \cdot (\text{Ad}_i \times \text{Campaign}_C) \\ &+ \sum_{k=2}^{51} \beta_{\text{State},k} \cdot \text{State}_k + \sum_{l=1}^{M} \beta_{\text{Set},l} \cdot \text{Set}_l\end{aligned} \tag{6}$$

Note: (1) all control ads are coded as the same 'control ad' in this analysis. (2) the explanation of all other terms in the equation can be found in Equations 1, 2, 3, 4.

(2) Likelihood Ratio Test This test compares the full model (including interaction terms) with a simplified model that excludes these interactions to assess their significance.

Results: Tables 22, 23, and 24 present the regression results for this part.

(1) CTR Across Campaign Types: Ads optimized for views versus those optimized for clicks yield lower click outcomes.

(2) Treatment Effects and Interaction Terms: 2023 Analysis: Out of 10 treatments, two had negative effects, five had positive effects, and three showed no significant effect. Regarding interaction terms, five were significantly negative, and one was significantly positive.

2024 Analysis: Out of 5 treatments, two had negative effects, and three had positive effects. One interaction term was significantly positive.

2023 and 2024 Combined Analysis:

(3) Performance by Ad Set: In 2023, Ad set C performed worse than Ad set A, while Ad set B's performance was statistically similar to that of Ad set A. In 2024, Ad set E performed similarly to Ad set D.

(4) Likelihood-Ratio Test: All analyses showed significant results with p-values indicating strong statistical significance. See Table 25 for details.

Interpretation Pooling across the ad sets, we find the same significant interactions remain when demographics are controlled. We confirm that the type of optimization matters for the results, and it's insufficient to control for the limited demographics that Facebook provides, suggesting that Facebook's algorithms must be targeting based on other characteristics in a way that affects the results of Facebook experiments.

6 Conclusion

The study indicates that any experiment conducted on Facebook cannot be interpreted as a randomized controlled trial due to the differential impact of the platform's targeting algorithms on different experimental conditions. We provide evidence that the results of any experiment conducted on Facebook could be influenced by the platform's targeting algorithm. Specifically, we document that different ways of running campaigns (specifically view/reach-optimized vs. click-optimized) on Facebook can lead to different targeting (proxied by demographic characteristics of audiences) and, consequently, to different experimental results. We found that algorithms play a significant role in shaping campaign outcomes, not only by influencing the limited audience demographics that Facebook provides but also through targeting on other unobserved dimensions.

We suggest that experiments conducted on Facebook require careful interpretation. Instead of assuming that a treatment ad's success means that, all else equal, the same people are more inclined to click on it than on a control ad, it's important to recognize the additional effect of the platform's algorithms. Facebook's algorithmic targeting means that a treatment ad might appear more successful simply because the algorithm is serving it to users who are more predisposed to click.

As a result, the campaign outcomes are shaped both by the audience's initial likelihood of clicking (due to Facebook's targeting) and the causal impact of seeing the ad. This makes it important to consider the role of the algorithm when interpreting experiment results, rather than solely attributing success to the ad itself.

This study contributes to the broader discourse on the generalizability of results derived from experiments conducted in algorithmic environments, such as Facebook. If the primary objective is to identify what works best within the Facebook, then experimenting directly on the platform will yield valuable insights into the most effective strategies (as long as the algorithm remains consistent). However, it's crucial to recognize that these results are limited in their applicability outside of Facebook, either to algorithmic environments with different algorithms or to non-algorithmic environments.

	ViewOpt	**ClickOpt**
18-24female	3650.0	102.0
18-24male	2862.0	99.0
25-34female	6488.0	433.0
25-34male	8345.0	156.0
35-44female	6130.0	800.0
35-44male	8596.0	360.0
45-54female	7141.0	1387.0
45-54male	9301.0	1178.0
55-64female	10174.0	3345.0
55-64male	11452.0	3756.0
65+female	17653.0	6297.0
65+male	16582.0	8790.0

Table 3: Reach Contingency Table - 2 Optimization Goals x Age-Gender (All Sets in 2023)

	control1	A1	A2	A3	A4
18-24female	32.0	132.0	80.0	208.0	20.0
18-24male	26.0	74.0	64.0	188.0	30.0
25-34female	69.0	239.0	126.0	484.0	48.0
25-34male	103.0	182.0	178.0	528.0	67.0
35-44female	77.0	183.0	92.0	384.0	58.0
35-44male	83.0	159.0	139.0	568.0	75.0
45-54female	91.0	175.0	94.0	392.0	88.0
45-54male	117.0	169.0	180.0	520.0	90.0
55-64female	189.0	261.0	212.0	550.0	144.0
55-64male	147.0	170.0	258.0	726.0	115.0
65+female	258.0	351.0	524.0	906.0	241.0
65+male	187.0	213.0	449.0	852.0	155.0

Table 4: Reach Contingency Table - Ads in Set A x Age-Gender (View Optimization)

	Control1	Control2	Control3
18-24female	32	23	27
18-24male	26	24	29
25-34female	69	63	54
25-34male	103	97	57
35-44female	77	68	47
35-44male	83	87	61
45-54female	91	99	82
45-54male	117	87	84
55-64female	189	148	136
55-64male	147	120	108
65+female	258	255	238
65+male	187	154	164

Table 5: Reach Contingency Table - Control Ads in 2023 x Age-Gender (View Optimization)

	control1	A1
18-24female	9.0	5.0
18-24male	6.0	4.0
25-34female	82.0	19.0
25-34male	36.0	17.0
35-44female	222.0	67.0
35-44male	77.0	44.0
45-54female	507.0	212.0
45-54male	127.0	85.0
55-64female	1267.0	610.0
55-64male	219.0	225.0
65+female	1633.0	1619.0
65+male	269.0	466.0

Table 6: Reach Contingency Table - One Control-Treatment Pair in Set A x Age-Gender (View Optimization)

	With Interactions	Without Interactions
ViewOpt	-2.61**	-3.01**
	(0.38)	(0.16)
A1	-0.05	-0.03
	(0.09)	(0.09)
A2	0.30**	0.30**
	(0.09)	(0.09)
A3	-0.26**	-0.33**
	(0.10)	(0.10)
A4	0.21**	0.21**
	(0.08)	(0.08)
A1 x ViewOpt	0.24	
	(0.47)	
A2 x ViewOpt	-0.23	
	(0.48)	
A3 x ViewOpt	-1.61**	
	(0.59)	
A4 x ViewOpt	-0.36	
	(0.59)	
Intercept	-2.68**	-2.67**
	(0.06)	(0.06)
Obs	33140	33140

Table 7: Baseline Model Results Comparison (Set A)

Note: The Likelihood Ratio Test results show a statistic of 16.30, with 4 degrees of freedom and a p-value of 0.0026.

	With Interactions	Without Interactions
ViewOpt	-3.04**	-4.38**
	(0.50)	(0.17)
B1	0.40**	0.37**
	(0.08)	(0.08)
B2	-0.87**	-0.88**
	(0.11)	(0.11)
B3	-0.12	-0.10
	(0.09)	(0.09)
B1 x ViewOpt	-2.02**	
	(0.58)	
B2 x ViewOpt	-1.24*	
	(0.59)	
B3 x ViewOpt	0.80	
	(0.61)	
Intercept	-2.70**	-2.69**
	(0.06)	(0.06)
Obs	70295	70295

Table 8: Baseline Model Results Comparison (Set B)

Note: The Likelihood Ratio Test show a statistic of 34.47, with 3 degrees of freedom and a p-value of 0.0000.

	With Interactions	Without Interactions
ViewOpt	-3.47**	-3.93**
	(0.71)	(0.13)
C1	0.33**	0.28**
	(0.09)	(0.09)
C2	0.39**	0.34**
	(0.09)	(0.08)
C3	0.06	0.15^{+}
	(0.09)	(0.08)
C1 x ViewOpt	-1.02	
	(0.74)	
C2 x ViewOpt	-1.24	
	(0.76)	
C3 x ViewOpt	2.07**	
	(0.73)	
Intercept	-2.84**	-2.83**
	(0.06)	(0.06)
Obs	64424	64424

Table 9: Baseline Model Results Comparison (Set C)

Note: The Likelihood Ratio Test results show a statistic of 118.17, with 3 degrees of freedom and a p-value of 0.0000.

	With Interactions	Without Interactions
ReachOpt	-4.86**	-4.96**
	(0.22)	(0.08)
D1	0.51**	0.50**
	(0.06)	(0.06)
D2	0.33**	0.32**
	(0.06)	(0.06)
D3	0.27**	0.27**
	(0.08)	(0.07)
D1 x ReachOpt	-0.20	
	(0.26)	
D2 x ReachOpt	-0.09	
	(0.26)	
D3 x ReachOpt	0.01	
	(0.29)	
Intercept	-3.00**	-2.99**
	(0.05)	(0.05)
Obs	381499	381499

Table 10: Baseline Model Results Comparison (Set D)

Note: The Likelihood Ratio Test results show a statistic of 1.14, with 3 degrees of freedom and a p-value of 0.7682.

	With Interactions	Without Interactions
ReachOpt	-4.63**	-4.55**
	(0.19)	(0.08)
E1	-0.14*	-0.11$^+$
	(0.07)	(0.07)
E2	0.19**	0.18**
	(0.07)	(0.06)
E1 x ReachOpt	0.31	
	(0.23)	
E2 x ReachOpt	-0.07	
	(0.23)	
Intercept	-2.98**	-2.99**
	(0.06)	(0.05)
Obs	329188	329188

Table 11: Baseline Model Results Comparison (Set E)

Note: The Likelihood Ratio Test results show a statistic of 4.55, with 2 degrees of freedom and a p-value of 0.1030.

Ad Set	**Likelihood Ratio Statistic**	**Degrees of Freedom**	**p-value**
Set A	16.30	4	0.0026
Set B	34.47	3	0.0000
Set C	118.17	3	0.0000
Set D	1.14	3	0.7682
Set E	4.55	2	0.1030

Table 12: Likelihood Ratio Statistics for Baseline Models

	Controlling for Region	Controlling for Age and Gender
ViewOpt	-2.61**	-2.43**
	(0.38)	(0.39)
A1	-0.04	-0.10
	(0.09)	(0.10)
A2	0.32**	0.23*
	(0.09)	(0.09)
A3	-0.24*	-0.23*
	(0.10)	(0.10)
A4	0.23**	0.19*
	(0.08)	(0.08)
A1 x ViewOpt	0.25	0.33
	(0.47)	(0.47)
A2 x ViewOpt	-0.24	-0.18
	(0.48)	(0.48)
A3 x ViewOpt	-1.62**	-1.57**
	(0.59)	(0.59)
A4 x ViewOpt	-0.37	-0.34
	(0.59)	(0.59)
Intercept	-2.93**	-3.25**
	(0.25)	(0.51)
Obs	33140	32707

Table 13: Controlling for Demographics - Model Results (Set A)

	Controlling for Region	Controlling for Age and Gender
ViewOpt	-3.04**	-2.87**
	(0.50)	(0.51)
B1	0.41**	0.48**
	(0.08)	(0.08)
B2	-0.87**	-0.89**
	(0.11)	(0.11)
B3	-0.12	-0.18*
	(0.09)	(0.09)
B1 x ViewOpt	-2.03**	-2.07**
	(0.58)	(0.58)
B2 x ViewOpt	-1.24*	-1.19*
	(0.60)	(0.60)
B3 x ViewOpt	0.80	0.83
	(0.61)	(0.61)
Intercept	-2.81**	-2.54**
	(0.25)	(0.40)
Obs	70295	69140

Table 14: Controlling for Demographics - Model Results (Set B)

	Controlling for Region	Controlling for Age and Gender
ViewOpt	-3.48**	-3.31**
	(0.71)	(0.71)
C1	0.32**	0.27**
	(0.09)	(0.09)
C2	0.39**	0.29**
	(0.09)	(0.09)
C3	0.06	0.01
	(0.09)	(0.09)
C1 x ViewOpt	-1.02	-0.94
	(0.74)	(0.75)
C2 x ViewOpt	-1.23	-1.08
	(0.76)	(0.76)
C3 x ViewOpt	2.07**	2.22**
	(0.73)	(0.74)
Intercept	-2.98**	-2.76**
	(0.23)	(0.38)
Obs	64424	63489

Table 15: Controlling for Demographics - Model Results (Set C)

	Controlling for Region	Controlling for Age and Gender
ReachOpt	-4.86**	-4.70**
	(0.22)	(0.22)
D1	0.51**	0.48**
	(0.06)	(0.06)
D2	0.32**	0.33**
	(0.06)	(0.06)
D3	0.26**	0.32**
	(0.08)	(0.08)
D1 x ReachOpt	-0.20	-0.16
	(0.26)	(0.26)
D2 x ReachOpt	-0.08	-0.09
	(0.26)	(0.26)
D3 x ReachOpt	0.02	-0.04
	(0.29)	(0.29)
Intercept	-2.81**	-2.30**
	(0.13)	(0.31)
Obs	381499	380878

Table 16: Controlling for Demographics - Model Results (Set D)

	Controlling for Region	Controlling for Age and Gender
ReachOpt	-4.63**	-4.54**
	(0.19)	(0.20)
E1	-0.14*	-0.14*
	(0.07)	(0.07)
E2	0.19**	0.19**
	(0.07)	(0.07)
E1 x ReachOpt	0.31	0.31
	(0.23)	(0.24)
E2 x ReachOpt	-0.07	-0.09
	(0.23)	(0.24)
Intercept	-2.89**	-3.38**
	(0.15)	(0.51)
Obs	329188	328426

Table 17: Controlling for Demographics - Model Results (Set E)

	With Interaction	Without Interaction
ViewOpt	-2.93**	-3.88**
	(0.28)	(0.09)
A3	-0.26**	-0.29**
	(0.10)	(0.10)
B1	0.41**	0.34**
	(0.08)	(0.08)
C1	0.34**	0.27**
	(0.09)	(0.08)
C2	0.40**	0.34**
	(0.08)	(0.08)
C3	0.07	0.15^{+}
	(0.09)	(0.08)
A2	0.29**	0.32**
	(0.09)	(0.09)
B2	-0.87**	-0.92**
	(0.11)	(0.11)
B3	-0.12	-0.10
	(0.09)	(0.09)
A4	0.21**	0.20*
	(0.08)	(0.08)
A1	-0.06	-0.02
	(0.09)	(0.09)
A3 x ViewOpt	-1.29*	
	(0.53)	
B1 x ViewOpt	-2.13**	
	(0.40)	
C1 x ViewOpt	-1.57**	
	(0.36)	
C2 x ViewOpt	-1.78**	
	(0.39)	
C3 x ViewOpt	1.53**	
	(0.34)	
A2 x ViewOpt	0.09	
	(0.40)	
B2 x ViewOpt	-1.34**	
	(0.42)	
B3 x ViewOpt	0.69	
	(0.44)	
A4 x ViewOpt	-0.04	
	(0.53)	
A1 x ViewOpt	0.56	
	(0.39)	
Intercept	-2.67**	-2.66**
	(0.06)	(0.06)
Obs	167859	167859

Table 18: Polling Across Ad Sets - Model Results (2023)

Note: The Likelihood Ratio Test results show a statistic of 202.61, with 10 degrees of freedom and a p-value of 0.0000.

	With Interaction	Without Interaction
ReachOpt	-4.73**	-4.78**
	(0.15)	(0.06)
D1	0.52**	0.50**
	(0.06)	(0.06)
D2	0.33**	0.32**
	(0.06)	(0.06)
D3	0.27**	0.27**
	(0.08)	(0.07)
E1	-0.15*	-0.11^{+}
	(0.07)	(0.07)
E2	0.18**	0.18**
	(0.06)	(0.06)
D1 x ReachOpt	-0.32	
	(0.20)	
D2 x ReachOpt	-0.21	
	(0.20)	
D3 x ReachOpt	-0.11	
	(0.24)	
E1 x ReachOpt	0.41*	
	(0.19)	
E2 x ReachOpt	0.03	
	(0.19)	
Intercept	-3.00**	-3.00**
	(0.05)	(0.05)
Obs	710687	710687

Table 19: Polling Across Ad Sets - Model Results (2024)

Note: The Likelihood Ratio Test results show a statistic of 18.03, with 5 degrees of freedom and a p-value of 0.0029.

	With Interaction	Without Interaction
ExposureOpt	-4.53**	-4.56**
	(0.13)	(0.05)
A3	-0.28**	-0.27**
	(0.10)	(0.10)
B1	0.39**	0.38**
	(0.08)	(0.08)
D1	0.54**	0.50**
	(0.06)	(0.06)
C1	0.32**	0.33**
	(0.09)	(0.08)
C2	0.39**	0.38**
	(0.08)	(0.08)
C3	0.06	0.15^{+}
	(0.09)	(0.08)
D2	0.35**	0.32**
	(0.06)	(0.06)
A2	0.28**	0.32**
	(0.09)	(0.09)
D3	0.29**	0.26**
	(0.08)	(0.07)
E1	-0.13^{+}	-0.11^{+}
	(0.07)	(0.07)
B2	-0.89**	-0.86**
	(0.11)	(0.11)
B3	-0.13	-0.10
	(0.09)	(0.09)
A4	0.19*	0.20*
	(0.08)	(0.08)
A1	-0.07	-0.01
	(0.09)	(0.09)
E2	0.20**	0.18**
	(0.06)	(0.06)
A3 x ExposureOpt	0.31	
	(0.47)	
B1 x ExposureOpt	-0.52^{+}	
	(0.31)	
D1 x ExposureOpt	-0.52**	
	(0.19)	
C1 x ExposureOpt	0.04	
	(0.26)	
C2 x ExposureOpt	-0.18	
	(0.30)	
C3 x ExposureOpt	3.13**	
	(0.23)	
D2 x ExposureOpt	-0.41*	
	(0.19)	
A2 x ExposureOpt	1.69**	
	(0.31)	
D3 x ExposureOpt	-0.31	
	(0.23)	
E1 x ExposureOpt	0.21	
	(0.18)	
B2 x ExposureOpt	0.26	
	(0.34)	
B3 x ExposureOpt	2.29**	
	(0.37)	
A4 x ExposureOpt	1.56**	
	(0.47)	
A1 x ExposureOpt	2.16**	
	(0.31)	
E2 x ExposureOpt	-0.17	
	(0.18)	
Intercept	-2.66**	-2.66**
	(0.06)	(0.06)
Obs	878546	878546

Table 20: Polling Across Ad Sets - Model Results (2023 and 2024)

Note: The Likelihood Ratio Test results show a statistic of 265.65, with 15 degrees of freedom and a p-value of 0.0000.

Ad Set	Likelihood Ratio Statistic	Degrees of Freedom	p-value
Sets in 2023	202.61	10	0.0000
Sets in 2024	18.03	5	0.0029
Sets in 2023 and 2024	265.65	15	0.0000

Table 21: Likelihood Ratio Statistics for Data Pooled Models

	Controlling for Region		Controlling for Age and Gender	
	With Interaction	No Interaction	With Interaction	No Interaction
ViewOpt	-2.93**	-3.88**	-2.78**	-3.68**
	(0.28)	(0.09)	(0.28)	(0.09)
A3	-0.25**	-0.28**	-0.24*	-0.26**
	(0.10)	(0.10)	(0.10)	(0.10)
B1	0.41**	0.35**	0.48**	0.41**
	(0.08)	(0.08)	(0.08)	(0.08)
C1	0.33**	0.27**	0.27**	0.21*
	(0.09)	(0.08)	(0.09)	(0.09)
C2	0.40**	0.34**	0.30**	0.25**
	(0.09)	(0.08)	(0.09)	(0.08)
C3	0.07	0.15^{+}	0.01	0.09
	(0.09)	(0.08)	(0.09)	(0.08)
A2	0.31**	0.33**	0.22*	0.24**
	(0.09)	(0.09)	(0.09)	(0.09)
B2	-0.86**	-0.91**	-0.90**	-0.95**
	(0.11)	(0.11)	(0.11)	(0.11)
B3	-0.11	-0.09	-0.18*	-0.16^{+}
	(0.09)	(0.09)	(0.09)	(0.09)
A4	0.22**	0.21**	0.18*	0.18*
	(0.08)	(0.08)	(0.08)	(0.08)
A1	-0.05	-0.01	-0.10	-0.06
	(0.09)	(0.09)	(0.09)	(0.09)
A3 x ViewOpt	-1.30*		-1.25*	
	(0.53)		(0.53)	
B1 x ViewOpt	-2.13**		-2.16**	
	(0.40)		(0.40)	
C1 x ViewOpt	-1.57**		-1.46**	
	(0.36)		(0.36)	
C2 x ViewOpt	-1.78**		-1.66**	
	(0.39)		(0.39)	
C3 x ViewOpt	1.53**		1.67**	
	(0.34)		(0.34)	
A2 x ViewOpt	0.08		0.15	
	(0.40)		(0.40)	
B2 x ViewOpt	-1.36**		-1.27**	
	(0.42)		(0.42)	
B3 x ViewOpt	0.68		0.75^{+}	
	(0.44)		(0.44)	
A4 x ViewOpt	-0.05		-0.02	
	(0.53)		(0.53)	
A1 x ViewOpt	0.56		0.64	
	(0.39)		(0.39)	
Intercept	-2.84**	-2.82**	-2.75**	-2.67**
	(0.15)	(0.15)	(0.24)	(0.24)
Obs	167859	167859	167763	167763

Table 22: Polling Across Ad Sets, Controlling for Demographics - Model Results (2023)

Note: For the ‘Controlling for Region’ Part, the Likelihood Ratio Test results show a statistic of 203.12, with 10 degrees of freedom and a p-value of 0.0000.
For the ‘Controlling for Age and Gender’ Part, the Likelihood Ratio Test results show a statistic of 208.41, with 10 degrees of freedom and a p-value of 0.0000.

	Controlling for Region		Controlling for Age and Gender	
	With Interaction	No Interaction	With Interaction	No Interaction
ReachOpt	-4.73^{**}	-4.77^{**}	-4.60^{**}	-4.64^{**}
	(0.15)	(0.06)	(0.15)	(0.06)
D1	0.52^{**}	0.50^{**}	0.50^{**}	0.48^{**}
	(0.06)	(0.06)	(0.06)	(0.06)
D2	0.33^{**}	0.32^{**}	0.34^{**}	0.33^{**}
	(0.06)	(0.06)	(0.06)	(0.06)
D3	0.27^{**}	0.26^{**}	0.31^{**}	0.30^{**}
	(0.08)	(0.07)	(0.08)	(0.07)
E1	-0.15^{*}	-0.12^{+}	-0.15^{*}	-0.12^{+}
	(0.07)	(0.07)	(0.07)	(0.07)
E2	0.18^{**}	0.18^{**}	0.18^{**}	0.18^{**}
	(0.06)	(0.06)	(0.06)	(0.06)
D1 x ReachOpt	-0.32		-0.28	
	(0.20)		(0.20)	
D2 x ReachOpt	-0.20		-0.20	
	(0.20)		(0.20)	
D3 x ReachOpt	-0.10		-0.14	
	(0.24)		(0.24)	
E1 x ReachOpt	0.42^{*}		0.42^{*}	
	(0.19)		(0.20)	
E2 x ReachOpt	0.03		0.01	
	(0.19)		(0.20)	
Intercept	-2.86^{**}	-2.86^{**}	-2.70^{**}	-2.72^{**}
	(0.10)	(0.10)	(0.26)	(0.26)
Obs	710687	710687	709304	709304

Table 23: Polling Across Ad Sets, Controlling for Demographics - Model Results (2024)

Note: For the ‘Controlling for Region’ Part, the Likelihood Ratio Test results show a statistic of 18.18, with 5 degrees of freedom and a p-value of 0.0027.
For the ‘Controlling for Age and Gender’ Part, the Likelihood Ratio Test results show a statistic of 16.44, with 5 degrees of freedom and a p-value of 0.0057.

	Controlling for Region		Controlling for Age and Gender	
	With Interaction	Without Interaction	With Interaction	Without Interaction
ExposureOpt	-4.53**	-4.56**	-4.33**	-4.37**
	(0.13)	(0.05)	(0.13)	(0.05)
A3	-0.28**	-0.27**	-0.27**	-0.26**
	(0.10)	(0.10)	(0.10)	(0.10)
B1	0.39**	0.38**	0.44**	0.42**
	(0.08)	(0.08)	(0.08)	(0.08)
D1	0.54**	0.50**	0.50**	0.47**
	(0.06)	(0.06)	(0.06)	(0.06)
C1	0.32**	0.32**	0.27**	0.27**
	(0.09)	(0.08)	(0.09)	(0.08)
C2	0.39**	0.38**	0.31**	0.30**
	(0.08)	(0.08)	(0.09)	(0.08)
C3	0.06	0.15^{+}	0.01	0.10
	(0.09)	(0.08)	(0.09)	(0.08)
D2	0.35**	0.32**	0.36**	0.33**
	(0.06)	(0.06)	(0.06)	(0.06)
A2	0.28**	0.33**	0.22*	0.26**
	(0.09)	(0.09)	(0.09)	(0.09)
D3	0.28**	0.26**	0.34**	0.31**
	(0.08)	(0.07)	(0.08)	(0.07)
E1	-0.14*	-0.12^{+}	-0.14*	-0.12^{+}
	(0.07)	(0.07)	(0.07)	(0.07)
B2	-0.88**	-0.86**	-0.91**	-0.90**
	(0.11)	(0.11)	(0.11)	(0.11)
B3	-0.13	-0.09	-0.19*	-0.16^{+}
	(0.09)	(0.09)	(0.09)	(0.09)
A4	0.20*	0.21**	0.17*	0.18*
	(0.08)	(0.08)	(0.08)	(0.08)
A1	-0.07	-0.01	-0.11	-0.05
	(0.09)	(0.09)	(0.09)	(0.09)
E2	0.20**	0.18**	0.19**	0.18**
	(0.06)	(0.06)	(0.06)	(0.06)
A3 x ExposureOpt	0.31		0.25	
	(0.47)		(0.47)	
B1 x ExposureOpt	-0.53^{+}		-0.64*	
	(0.31)		(0.31)	
D1 x ExposureOpt	-0.52**		-0.47*	
	(0.19)		(0.19)	
C1 x ExposureOpt	0.04		0.03	
	(0.26)		(0.26)	
C2 x ExposureOpt	-0.18		-0.18	
	(0.30)		(0.30)	
C3 x ExposureOpt	3.13**		3.14**	
	(0.23)		(0.23)	
D2 x ExposureOpt	-0.40*		-0.40*	
	(0.19)		(0.19)	
A2 x ExposureOpt	1.68**		1.64**	
	(0.31)		(0.31)	
D3 x ExposureOpt	-0.30		-0.36	
	(0.23)		(0.23)	
E1 x ExposureOpt	0.21		0.22	
	(0.18)		(0.19)	
B2 x ExposureOpt	0.25		0.22	
	(0.34)		(0.34)	
B3 x ExposureOpt	2.29**		2.24**	
	(0.37)		(0.37)	
A4 x ExposureOpt	1.56**		1.49**	
	(0.47)		(0.47)	
A1 x ExposureOpt	2.17**		2.13**	
	(0.31)		(0.31)	
E2 x ExposureOpt	-0.17		-0.19	
	(0.18)		(0.19)	
Intercept	-2.62**	-2.62**	-2.59**	-2.47**
	(0.10)	(0.10)	(0.18)	(0.18)
Obs	878546	878546	877067	877067

Table 24: Polling Across Ad Sets, Controlling for Demographics - Model Results (2023 and 2024)

Note: For the ‘Controlling for Region’ Part, the Likelihood Ratio Test results show a statistic of 265.34, with 15 degrees of freedom and a p-value of 0.0000.
For the ‘Controlling for Age and Gender’ Part, the Likelihood Ratio Test results show a statistic of 262.71, with 15 degrees of freedom and a p-value of 0.0000.

Ad Set	**Likelihood Ratio Statistic**	**Degrees of Freedom**	**p-value**
Sets in 2023, Region Control	203.12	10	0.0000
Sets in 2023, Age-Gender Control	208.41	10	0.0000
Sets in 2024, Region Control	18.18	5	0.0027
Sets in 2024, Age-Gender Control	16.44	5	0.0057
Sets in 2023 and 2024, Region Control	265.34	15	0.0000
Sets in 2023 and 2024, Age-Gender Control	262.71	15	0.0000

Table 25: Likelihood Ratio Statistics for Data Pooled Models with Controls

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction*, *3*(CSCW), 1–30.

Eckles, D., Gordon, B. R., & Johnson, G. A. (2018). Field studies of psychologically targeted ads face threats to internal validity. *Proceedings of the National Academy of Sciences*, *115*(23), E5254–E5255.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist*, *70*(6), 543.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, *114*(48), 12714–12719.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2018). Reply to eckles et al.: Facebook's optimization algorithms are highly unlikely to explain the effects of psychological targeting. *Proceedings of the National Academy of Sciences*, *115*(23), E5256–E5257.

Meta for Business. (2018). Making split testing easier and more accessible for all advertisers [Accessed: 2024-05-15]. https://www.facebook.com/business/news/making-split-testing-easier-and-more-accessible-for-all-advertisers/

Meta for Business. (2024). About a/b testing [Accessed: 2024-05-15]. https://www.facebook.com/business/help/1738164643098669?id=445653312788501

Orazi, D. C., & Johnston, A. C. (2020). Running field experiments using facebook split test. *Journal of Business Research*, *118*, 189–198.