7037 Group M Final Project

Xia Yi 3036346450

Li Guiquan 3036347131

Huang Haowen 3036345298

Feng Jiayi 3036347234

Wu Zhuoya 3036345834

1. Form Hypotheses

H_0: Showing a friend's like in the ad has no impact on click-through rate.

H_1: Showing a friend's like increases the click-through rate.

2. Sanity Check

In sanity check, we conduct three different type of check to ensure the scientific nature and credibility of the subsequent analysis results.

In SRM check, we use Chi-square homogeneity check and t-test check to check if the control/treatment group split aligns with expectations. In Covariate balance check, we use t-test and Mann-Whitney U test to ensure comparability between groups. In A/A test, we run the experiment without any intervention to verify whether the platform can generate results with no differences between groups.

The experiment has passed the sanity check, indicating that the randomization mechanism of the experiment is effective and the data quality is reliable.

3. Power Analysis

In this part of the analysis, we evaluated whether our experiment had sufficient statistical power to detect the treatment effect on click-through rate, particularly focusing on the smallest observed lift across product categories.

In order to do this, we first extracted the minimum effect size (ΔCTR = 0.0068) from the category-level CTR summary. This value corresponds to the smallest observed difference between the Treatment and Control groups—specifically from the Cosmetics category. We then used the Control group's baseline CTR (p_1 = 0.0677) to calculate the effect size in terms of Cohen's h.

To quantify the statistical power, we applied the NormalIndPower class from the statsmodels package and conducted two key analyses:(1).Post-hoc power analysis: Given our actual sample size of 100,000 observations per group and a significance level of α = 0.05, we found that the power to detect a 0.0068 lift was approximately 1.000. This means that our study is virtually guaranteed to detect such an effect if it exists. (2).A-priori sample size calculation: To achieve 80% power for detecting the same 0.0068 lift, we would only need approximately 25,756 samples per group. We also computed required sample sizes for a range of minimum detectable effects (MDEs) from 0.002 to 0.015 and visualized the results using a power curve.

As shown in the power curve, smaller effects require substantially larger sample sizes—for instance, detecting a 0.4 percentage point lift would require over 50,000 users per group, while a 1.5 percentage point lift would require fewer than 5,000.

In conclusion, the power analysis confirms that our current sample size is more than sufficient to detect even the smallest observed treatment effects, with post-hoc power approaching 1.000. This ensures that any non-significant findings in subsequent analyses are unlikely to be due to insufficient data. Moreover, the power curve provides a practical reference for future experiments by linking detectable effect sizes to required sample sizes, thereby supporting more efficient and confident experimental design.

4. Compare test

We need to calculate the CTR mean and standard deviation for control and treatment groups. And we use t-test to check if the difference is significant and use levins test to check if the variance is equal.

According to the code results:

- Control Group (0): CTR = 6.66% (SD = 0.249)
- Treatment Group (1): CTR = 7.45% (SD = 0.263)
- T-test: t=15.43, p<0.001
- Levene's Test: F=238.10, p<0.001

We can summarize the results as follows:

（1）Statistically Significant Improvement: The treatment group (showing 1 organic like) achieves a 0.79% absolute increase in CTR (11.8% relative increase) compared to the control group. The extremely low p-value (p<0.001) confirms this difference is not due to chance. For context, even small CTR improvements in large-scale platforms like WeChat Moments Ads can translate to millions of incremental clicks.

（2）Unequal Variance: The Levene's test (p<0.001) rejects equal variance between groups. The treatment group exhibits higher variance in CTR (σ2=0.069) than the control (σ2=0.062), suggesting heterogeneity in user responses to social proof. Some users in the treatment group may be more influenced by the displayed like, while others remain indifferent.

Also, to make testing more reliable, we can use bootstrap method to calculate the confidence interval for CTR.

According to the code results, we can find: the 95% CI does not include zero, reinforcing the t-test conclusion (p<0.001) that the treatment effect is statistically significant. Also, the narrow CI (0.0069–0.0089) suggests high certainty that the true CTR lift lies between 0.6% and 0.9%. This is critical for quantifying the business impact.

To reduce estimation variance and improve sensitivity in detecting differences in click-through rates (CTR) between treatment and control groups, we applied the following analytical strategies:

CUPED Adjustment (Single & Multiple Covariates): We performed covariance pre-adjustment using behavioral variables such as user_sns_like_cnt, user_sns_comment_cnt, user_degree, and real_like_cnt.Under single-variable CUPED, CTRs were 6.66% (Control) and 7.45% (Treatment), with minimal variance reduction—indicating that the covariate alone is weak. Multivariate CUPED slightly improved variance reduction, but sensitivity remains largely driven by large sample size.

Stratified Analysis: We stratified users by discrete variables like gender, city, and category, then calculated within-stratum CTRs and combined them via weighted averages.Category stratification led to the greatest sensitivity gain, reducing SE by 2.14% (Control) and 2.98% (Treatment), confirming that product category explains significant CTR heterogeneity.Gender and City had negligible impact on variance, indicating limited stratification value.

Covariate-Adjusted Linear Regression: A linear model controlling for demographic, behavioral, and ad-level variables was estimated.The treatment variable has a significant positive coefficient (0.0079, p < 0.001), confirming a strong treatment effect. Brand-related and experience-good ads significantly reduce CTR. CTR significantly declines in week 2 and 3 relative to week 1.

5. Analyze the Heterogeneous Treatment Effects (HTE)

In this analysis, we first loaded and cleaned the data, lowercased all column names, and constructed key variables: a binary treatment indicator (treat), product-type flags for status and experience goods, a standardized friend-relative-status score (rel_status_z, defined as the z-score of friend_degree minus user_degree) to capture how a friend's social standing compares to the user's, and a standardized friend-like count (real_like_z, the z-score of real_like_cnt) to measure the strength of social-proof at first ad impression. We also converted gender, week, city, and other categorical fields to factor data types.

Examining the average treatment effect (ATE), we found that the click-through rate rose from

6.66% in the control group to 7.45% in the treatment group—an absolute lift of 0.79 percentage points ($p < 0.001$). Next, we tested for heterogeneous treatment effects (HTE) by adding interaction terms in a logistic regression: treatment × status, treatment × experience, treatment × rel_status_z, and treatment × real_like_z, as well as three-way interactions combining product type and friend attributes. None of these interaction coefficients reached statistical significance, indicating that the treatment uplift is consistent across both status- and experience-type products and does not vary meaningfully with a friend's relative status or the number of friends who have already "liked" the ad.

Finally, we reran the main specification using clustered standard errors at the ad level (adid). While clustering widened standard errors on friend-level covariates—rendering rel_status_z and real_like_z non-significant— the overall treatment effect remained highly significant, and the absence of product-type heterogeneity persisted. In summary, revealing friends' likes in the treatment condition produces a robust, positive lift in click-through rate, with no evidence of differential impact across product categories or friend characteristics.

6. Summarize your results

This experiment aimed to evaluate the impact of displaying a friend's organic like in WeChat Moments Ads on user click-through behavior. To ensure the reliability and scientific validity of the results, the analysis followed a rigorous and structured process across multiple key phases. The study began with the formulation of a clear hypothesis: the primary hypothesis (H_1) proposed that showing a friend's like within an ad would increase the click-through rate (CTR), compared to a control condition in which likes were not shown. To validate the experimental setup, a comprehensive set of sanity checks was conducted. These included a Sample Ratio Mismatch (SRM) check, covariate balance tests (such as t-tests and Mann-Whitney U tests), and an A/A test to confirm the randomization procedure. The experiment successfully passed all these checks, confirming that both the group assignment mechanism and the quality of the data were suitable for causal inference.

Subsequently, a power analysis was carried out to determine whether the experiment had sufficient statistical power to detect meaningful treatment effects. With 100,000 observations per group, the analysis revealed that the study had extremely high power (approximately 1.000) to detect even the smallest observed CTR lift (0.68%), which was particularly important for interpreting null findings in the later subgroup analyses. The core result of the study, the average treatment effect (ATE), showed that the treatment group achieved a CTR of 7.45%, compared to 6.66% in the control group—an absolute lift of 0.79 percentage points, representing an 11.8% relative increase. This difference was highly statistically significant, with a t-statistic of 15.43 and a p-value less than 0.001. Bootstrap methods were used to reinforce the robustness of this finding, yielding a narrow and informative 95% confidence interval of [0.0069, 0.0089], indicating both statistical precision and practical significance.

Beyond average effects, the analysis also explored differences in response variability and ways to enhance test sensitivity. Levene's test showed that the treatment group had significantly higher variance in CTR, suggesting that users differ in how they respond to visible social proof. To address variance and improve the detection of treatment effects, techniques such as CUPED (using behavioral covariates like user activity and network size) and stratified analysis (particularly by product category) were applied. These approaches led to modest improvements in sensitivity, and a covariate-adjusted linear regression further confirmed the treatment effect (coefficient = 0.0079, $p < 0.001$), while also revealing that some ad and user-level characteristics (e.g., brand identity, experience goods, and week of exposure) significantly influenced CTR.

The analysis of heterogeneous treatment effects (HTE) was designed to determine whether certain product types or friend attributes moderated the treatment effect. Interaction terms for treatment with status-type goods, experience-type goods, friend-relative-status (rel_status_z), and the number of likes (real_like_z) were tested in logistic regression models. None of these interactions reached statistical significance, indicating that the treatment effect was stable and consistent across different subgroups. Importantly, these conclusions held even when clustered standard errors were applied at the ad level (adid), reinforcing the robustness and generalizability of the results.

In summary, the experiment provides strong and reliable evidence that displaying a single friend's like within WeChat Moments Ads significantly boosts click-through rates. This effect is robust, statistically significant, and consistent across user demographics, product categories, and friend-level characteristics. The findings suggest that social proof in the form of one friend's like is a simple, effective, and scalable strategy for enhancing ad engagement across the platform.

7. Inform product strategies for WeChat Moments Ads based on your results

Based on the robust findings from this experiment, several product strategy recommendations can be made to enhance the effectiveness of WeChat Moments Ads. First and foremost, the display of a single friend's like should be adopted as a default feature in ad presentation, particularly during the user's first ad impression. The experiment demonstrated a consistent and significant uplift in CTR when this element of social proof was shown, and this effect was stable across various product categories and user segments. Because the treatment effect does not depend on product type (e.g., status or experience goods) or friend-level attributes (such as relative status or popularity), there is no need to overcomplicate the logic behind which friend's like to display or under what conditions to show it. This simplicity offers scalability and ease of implementation without sacrificing performance.

Second, while no significant heterogeneity was found in the treatment effect across subgroups, the baseline CTR did vary by product category. For instance, certain categories like cosmetics showed lower CTRs overall. This implies that category-specific optimization—such as adjusting creative design, budget allocation, or engagement triggers—could further enhance performance, even if the treatment lift is similar across categories. Category-based targeting should therefore focus more on amplifying base engagement, while continuing to apply the same social proof logic uniformly.

Third, given that Levene's test indicated greater variance in the treatment group, it's likely that users respond to social proof differently—some are highly influenced by seeing a friend's like, while others may be indifferent. This opens up the opportunity for future segmentation-based strategies or A/B testing to identify and prioritize "high-response" subgroups, such as users with stronger social engagement history, higher network connectivity, or behavioral similarity to the liking friend.

Fourth, from a testing and experimentation perspective, the power analysis revealed that a relatively small sample size (around 25,000 users per group) would have been sufficient to detect even subtle CTR lifts like 0.0068 with 80% power. This insight can be used to design more efficient future experiments, especially for micro-optimizations or UI variations. In addition, bootstrap-based confidence intervals and stratified analysis by category were shown to increase sensitivity without needing extra data, offering practical tools for making A/B tests more reliable and cost-effective.

Finally, while the display of a single like was effective, the success of this intervention invites further iteration. Future research might explore the incremental impact of showing multiple friend likes, dynamic placement of social proof within the ad creative, or timing-related effects (e.g., whether early or late ad views benefit more from social proof). Such refinements could build on the strong baseline established by this experiment.

In conclusion, the use of social proof through friend likes presents a lightweight, impactful, and broadly applicable enhancement to WeChat Moments Ads. The consistent effect across users and products simplifies deployment, while insights from variance and power analysis can guide future iterations toward even more personalized, efficient, and engaging ad experiences.