Summary of AB Test of Udacity Website

Jenny Hung

1. Experiment Overview

Currently, Udacity courses have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

1.1 Metric Choice

In conducting the experiment, our goal is to test for whether or not the screening question prompts the student to enroll in free trial only after careful assessment of one’s ability to commit substantial study time to a course. The hypothesis is that this might reduce the number of students who do not go pass the trial, without significantly reducing the number of students who do.

Invariant metrics: Number of cookies, number of clicks and the click-through probability. An invariant metric should not change across experimental and control groups. Because the screener pops-up after clicking on the ’start free trial’ button, the number of pageviews, clicks, and the click-through-probability should remain unchanged during the experiment.

Evaluation metrics: Gross conversion, retention, and net conversion. Evaluation metrics are expected to change over the course of the experiment. Gross conversion, retention, and net conversions are all ratios which incorporates user ids (whose numbers are hypothesized to change) and the invariant cookie counts. Note that user id is not selected as an evaluation metric for the fact that it is measure of counts – it by itself is poor measure of the experimental effect as the change in counts could just be a natural variation, rather than due to the experimental effect. By using ratios listed above, we are better able to measure the effect of the screener.

In order to launch the experiment, we must observe the following: Gross Conversion is lower, since if the experiment is effective, the students who could not invest the required study time would have been screened out, resulting in a lower conversion rates. Retention would have been higher if the experiment is effective since those have enrolled are less likely to drop. However, we would also need to see Net Conversion remain constant, since if the experiment is effective, the number of user id that go pass the trial is expected to remain relatively constant.

1.2 Measuring Standard Deviation

Standard deviation for the evaluation metrics:

Evaluation Metrics	Standard Deviation
Gross conversion	0.0202
Retention	0.0549
Net conversion	0.0156

The above are the analytic standard deviations for the evaluation metrics. However, had we computed the empirical standard deviations for Retention, it would have been much more variable compared to the reported analytic one – this is because the unit of diversion (cookies) differs from the unit of analysis (which is the denominator of the Retention metric, user-ids). However, the empirical standard deviation for Gross Conversion and Net Conversion would have been at comparable levels compared to the analytic standard deviations, given that the unit of diversion is the same as the unit of analysis in this case.

1.3 Sizing

Number of Samples vs. Power

The number of pageviews needed to obtain the desirable power is 685,325. Bonferroni correction was not used in this case given that it would have resulted in a much too conservative pageviews.

Note that the required pageview count of 685,325 is arrived after we exclude the Retention as an evaluation metric. We had initially used Retention to compute the required pageview. However, this resulted in an unacceptable duration of 119 days while diverting 100% of web traffic to the experiment. After close inspection, it was decided that we can remove retention as an evaluation metric while maintaining sufficient coverage, given the fact that Retention measures the rate at which people remain in enrollment after the 14-day trial period, and Net Conversion measures the rate that a cookie turns into a user Id AND stays in enrollment after the trial period. With this decision, we are safe to use the combination of Net Conversion and Gross Conversion as evaluation metrics due to the high level of correlation.

Duration and Exposure

In considering the required pageviews to determine the duration, we have considered the exposure to risk for Udacity. Since Udacity is neither collecting additional sensitive or confidential information, nor subjecting any students to harm or embarrassment, the exposure is low. Therefore we can recommend diverting 100% of the web traffic to the experiment in order to shorten the duration for the experiment.

2. Experiment Analysis

2.1 Sanity Checks

Summary of confidence interval as sanity check conclusions are reported below:

Metric	Lower Bound	Upper Bound	Observed	Sanity Check
Number of cookies	0.4988	0.5012	0.5006	Pass
Number of clicks	0.4959	0.5041	0.5005	Pass
Click-through probability	-0.0013	0.0013	0.0001	Pass

2.2 Result Analysis

Effect Size Tests

Summary of confidence interval and conclusion of significance for the evaluation metrics are reported below:

Metric	Lower Bound	Upper Bound	Statistical Significance	Practical Significance
Gross Conversion	-0.0291	-0.0120	Yes	Yes
Net Conversion	-0.0116	0.0019	No	No

Sign Tests

Summary of p-value and conclusion of sign test for the evaluation metrics are reported below:

Metric	P-value	Sign Test
Gross Conversion	0.0026	Significant
Net Conversion	0.6776	Not significant

2.3 Summary

Bonferroni correction was not used in my analysis. The reason for this was three-fold: 1. Bonferroni correction would have resulted in a much more conservative number of pageviews than is necessary. 2. Bonferroni correction is more aptly applied in situations where multiple statistical tests, especially where attaining statistical significance in one test results in attaining statistical significance overall. However, in this analysis, I have taken the approach of making sure that the experiment is recommended to launch after statistical as well as practical significance is obtained on BOTH evaluation metrics. 3. Given that the evaluation metrics are highly correlated, we are able to choose the Net Conversion as a basis to compute the number of pageviews required. Using this metric in this planning computation results in a much lower pageviews without sacrificing power.

3. Recommendation

Based on the analysis, I would not recommend Udacity to launch the new practice. This is based on the analysis of the evaluation metrics of Gross Conversion and Net Conversion. Since the Gross Conversion has demonstrated both the statistical significance as well as practice significance, it indicates that the screener is effective in reducing the number of students who drops out during the trial period. However, we are not able to obtain statistical significance on the Net Conversion ratio, although it does attain the practical significance level. Due to the design of the experiment and the requirement that both of the evaluation metrics be significant, we are not able to recommend launching of the experiment.

4. Follow-Up Experiment

A follow-up, or alternative experiment would be of interest, given the risky nature of the current experiment considered. This alternative experiment involves restricting the scope of the pageviews originated from North America, using a cookie as the unit of diversion, and the same evaluation metrics of Gross Conversion, Retention and Net Conversion metrics.

I also recommend the alternative experiment to also have the identical screener as proposed in the original experiment. In other words, the only difference between the original and the alterative experiments are the type of web traffic that is under-study. Using only the North American cookies will means that less web traffic overall is impacted by this experiment, which makes it less risky. In addition, having North American-sourced cookies is a proxy to North American users and this more homogeneous pool of sample also means that there will likely be a reduced variability in the sample. We can also adjust the practice significant level in the alternative design. These factors will all contribute to reducing the risk level for Udacity and produce a more conclusive experiment result.