A/B Testing Project - Udacity Free Trial Screener

Experiment Overview

The Udacity course home pages have two options: "start free trial" and "access course materials." Clicking "start free trial" prompts the user to enter their credit card information, subsequently enrolling them in a 14 day free trial of the course, after which they are automatically charged. Users who click "access course materials" will be able to view course content but receive no coaching support, verified certificate, or project feedback.

For this experiment Udacity tested a change wherein those users who clicked "start free trial" were asked how much time they were willing to devote to the course. Users choosing 5 or more hours per week would be taken through the checkout process as usual. For users indicating fewer than 5 hours per week a message would appear indicating the need for a greater time commitment to enable success and suggesting they might like to access the free content. At this point the student would have the option to continue enrolling in the free trial or access the course materials for free.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

Experiment Design

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.

Metric Choice

Invariant Metrics

Invariant metrics should not change across experiment and control groups, which can provide a way to double check the integrity of the experiment design after the experiment was conducted. In this experiment, the screener change shows up after clicking on the “start free trial” button, thus, the number of pageviews, clicks, and the click-through-probability are expected to remain the same for both control and experiment groups. Therefore, the invariance metrics chosen for this experiment are:

- Number of cookies (*Number of unique cookies to view the course overview page*)

- Number of clicks (*Number of unique cookies to click the start free trial button*)

Evaluation Metrics

Evaluation metrics are dependent on the experiment, and are expected to change over the course of the experiment. Each evaluation metric is associated with a minimum difference (dmin) that must be observed for consideration in the decision to launch the experiment. In this experiment, anything after the screener change shows up, number of user-id's, gross conversion, retention, and net conversion could be affected for control and experiment groups. Therefore, the evaluation metrics chosen for this experiment are:

- Gross conversion (*Number of users who enrolled in the free trial/Number of users who clicked the "Start Free Trial" button*): It is directly dependent on the experiment and could measure could measure whether or not the screener had an effect on enrollment. We expect that the value will be lower in the experiment group because those users who are likely to drop during the 14-day trail (are not able to commit more than 5 hours per week) would be filtered by the screener. Vice versa, for the control group, users won't see any pop-up message so they will enroll without any consideration of the number of hours they can commit per week. Therefore, the gross conversion in the control group is expected to be higher than in the experiment group.
- Retention (*Number of user-ids to remain enrolled for 14 days trial period and make their first payment/Number of users who enrolled in the free trial*): It is directly dependent on the experiment and could measure could measure whether or not the screener change had an effect on the 14-day retention rate. We expect the value will be higher in the experiment group because majority of the users are those who can commit 5 hours per week and are more likely to remain enrolled after 14-day period and start their first initial payment. But for the control group, the users enroll for the course without considering the time availability commitment, which might cause lower retention rate.
- Net conversion (*Number of user-ids remained enrolled for 14 days trial and at least make their first payment/Number of users clicked the Start Free Trial button*): It is directly dependent on the experiment and could measure could measure whether or not the screener change had an effect on the first payment completion rate. It would be great if this value increases after the experiment. But considering the expected decrease in the total number of users enrolling the 14-day free trail, we expect this values does not decrease significantly after the experiment.

Unused Metrics

- Number of User-ID (*Number of users who enroll in the free trial*): This is not an appropriate invariant metric because user-ids are tracked only after student enrolling in the free trial, so the number of user-ids might be different between the control and experiment groups. It is also not a good evaluation metric because this value does not have a denominator so it cannot be normalized.

- Click-through-probability (*Number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page*): This is a good invariant metric since the clicks are occurred before the users see the experiment, therefore it is does not depend on our test which is an excellent invariant metric. However, the number of cookies & number of clicks are already sufficient to use as invariant metric, thus this metric may not be redundant for analysis.

Measuring Standard Deviation

The number of clicks and enrollments follows a binomial distribution, and by the central limit theorem, the distribution of the three rates (gross conversion, retention, and net conversion) is Gaussian. The standard deviation of these normally distributed rates can be calculated using the formula:

$$\sigma = \sqrt{\frac{p(1-p)}{n}}$$

Given the daily sample of 5000 cookies, the number of clicks and enrollments can be calculated using the baseline values:

Number of click: $\frac{5000 \times 3200}{40000} = 400$

Number of enrollment: $\frac{5000 \times 660}{40000} = 82.5$

Table1. Analytical Estimate of Standard Deviation

Evaluation Metric	Baseline Value	Standard Deviation
Gross conversion	0.20625	0.0202
Retention	0.53	0.0549
Net conversion	0.1093125	0.0156

Both gross conversion and net conversion using number of cookies as denominator, which is also unit of diversion. Here, the unit of diversion and unit of analysis are the same, which indicate the analytical estimate would be comparable to the empirical standard deviation.

For retention, the denominator is "number of users enrolled the courseware", which is not similar as unit of diversion. Thus, the empirical variability may be different from the analytical estimate, thus we perform both an analytical and empirical estimate for this metric.

Sizing

Number of Samples vs. Power

Given the type I error rate of α equals 0.05, type II error β equals 0.20, and the minimum detectable effect for each evaluation metric, the sample size required to power the experiment appropriately can be calculated using Evan Miller. Then, the total number of pageviews can be calculated using the given unit to pageview ratio.

Click/pageview ratio = 3200/40000 =0.08
Enrollment/pageview ratio = 660/40000 = .0165

Table2. Results of sample size calculation

Evaluation Metric	Baseline Value	Minimum Detectable Effect	Sample size	Unit/pageview ratio	Total number of pageviews
Gross conversion	0.20625	0.0202	25838	0.08	645,950
Retention	0.53	0.0549	39115	0.0165	4,741,213
Net conversion	0.1093125	0.0156	27413	0.08	685,325

Based on the results in table 2, a total of 4,741,213 pageviews is required to conduct the experiment.

Duration vs. Exposure

With daily pageview baseline value of 40000, the number of pageview for retention would need about 119 days, even if we divert 100% of traffic. It is unreasonably long for an A/B testing experiment. Therefore, I eliminate retention as the evaluation metric. The total number of required pageviews is decreased to 685,325. Considering that this is not a risky experiment as the change is small and it won't cause too much trouble in the overall business, I choose to direct 70% of the traffic (40000*0.7 =28000) to the experiment. Thus, it would takes approximately 25 days (685,325/28000 =25) to run the experiment.

Experiment Analysis

Sanity Checks

Having conducted the experiment, each of the invariant metrics needs double-check whether the underlying assumptions are being met. Cookies and clicks are expected to be divided evenly between the control and experimental groups. Using an expected rate of diversion of 0.5, the standard deviation can be calculated and a 95% confidence interval can be constructed around the expected value.

According to the results of table 3, both invariant metrics, cookie and click, pass the sanity check since their observed values are within 95% confidence interval.

Table3. Results of sanity checks

Invariant Metric	Expected value	Observed value	CI Lower Bound	CI Upper Bound	Results
Number of cookies	0.50	0.499360	0.496268	0.503732	Pass
Number of clicks	0.50	0.499533	0.486982	0.513018	Pass

Result Analysis

Effect Size Tests

For each evaluation metric, statistical and practical significance (whether or not the size of the effect is relevant from a business standpoint) should be tested. The minimum detectable effect is the smallest difference that we will accept between experimental and control groups in order to be practically significant.

Using the data collected, we calculate the rate in experimental and control groups for each evaluation metric (gross conversion, net conversion), and then define a new variable that is the difference between the rates (experiment - control). Using this newly defined variable, we construct a confidence interval which will then set a range for the expected difference.

Table4. Results of effect size tests

Evaluation Metric	d-min	Observed diff	CI Lower Bound	CI Upper Bound	Results
Gross conversion	0.01	-0.0206	-0.0291	-0.0120	Statistically and practically significant
Net conversion	0.0075	-0.0049	-0.0116	0.0019	Neither statistically nor practically significant

Since 95% confidence interval does not include zero and the minimum detectable effect value, gross conversion is both statistically and practically significant. In terms of net conversion, the 95% confidence interval includes zero and the minimum detectable effect value, indicating neither statistically nor practically significant.

Sign Tests

A binomial sign test will be conducted to further test each of the evaluation metrics. Each day of the experiment is evaluated to see if there is a positive or negative difference across groups (experimental - control). Each positive difference is counted as a success, and each negative

difference as a failure. Comparing the resulting p-values for each metric to determine significance.

Table5. Results of sign tests

Evaluation Metric	# of success	# of trails	Probability	Two-tails p-value	Results
Gross conversion	4	23	0.5	0.0025	Statistically significant
Net conversion	10	23	0.5	0.6776	statistically nonsignificant

According to the results of table 5, gross conversion rate has 4 of 23 successes for a two-tailed p-value of 0.0026 indicating statistical significance of gross conversion. Net conversion has 10 of 23 successes and a two-tailed p-value of 0.6776 indicating that net conversion is not statistically significant. Both are consistent with the hypothesis test results.

Summary

In this experiment, potential Udacity users were diverted by cookie into either control group or experiment group. After clicking "start free trail" button on the home page, users in the experiment group were asked how much time they are willing to devote to the course, while users in the control group were not. This experiment was designed to determine whether filtering users as a function of study time commitment would improve the overall user experience and improve coaches' capacity to support users who are likely to complete the course, without significantly reducing the number of students who continue past the free trial. Number of Cookies, Number of clicks on "start free trial" button were chosen as invariant metrics while Gross Conversion (enrollment/cookie) and Net Conversion (payments/cookie) were chosen as evaluation metrics. The null hypothesis is that there is no significant difference in the two evaluation metrics between control and experiment groups. In order to launch the experiment, the null hypothesis must be rejected for all evaluation metrics, as well as the differences between two groups should larger than the minimum detectable effect for each evaluation metric. Because the evaluation metrics in the experiment have high correlation and all evaluation metrics must have statically significant differences, thus the Bonferroni correction will be too conservative and will not be used during the analysis phase.

The sanity test results revealed, for the two invariant metrics, the expected equal distribution of cookies into the control and experiment group at the 95% confidence interval. In terms of the effect size hypothesis tests, the difference in Gross Conversion between the control and experiment group was both statistically and practically significant, and the null hypothesis was rejected. However, the difference in Net Conversion was neither statically nor practically significant at the 95% CI.

Recommendation

Based on the analysis results, I do not recommend to launch this experiment. The reasons are as follows: 1) although Gross Conversion turned out to be negative and practically significant, Net Conversion results are both statistically and practically insignificant, which did not meet the acceptance criteria that the null hypothesis must be rejected for all evaluation metrics. 2) The 95% confidence interval of Net Conversion does include the negative number of the practical significance boundary, which suggests the risk of hurting Udacity's business - decrease the revenue.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

In order to reduce the drop-out rate before the end of the 14 day trial period, a follow-up experiment is recommended. A mini-project that covers the prerequisite skills and the basic knowledge of the chosen course is designed for users to complete in the 14 day trail period. The goal of this mini-project is to help users to self-evaluate their competence and build their confidence for completing the course. We expect the users in the experiment group who conduct this mini-project have a lower drop-out rate before the end of the 14 day trail period than users in the control group.

The unit o diversion would be user-id because user once enrolled in the courses can be tracked by their user-ids.

Invariant metrics:

- Number of user-ids (Number of unique user-id to click the start free trial button)

Evaluation metrics:

- Retention (Number of user-ids to remain enrolled for 14 days trial period and make their first payment/Number of users who enrolled in the free trial)