A/B Testing – Free Trial Scanner

Amandeep Mander

Introduction

In this experiment, Udacity tested the change where if the student clicked “Start free trial” on the home page, they were asked about their availability to commit to the course. If the student indicated that they had fewer than 5 hours per week, a suggestion would be made to instead access the course materials for free. The student would then be given the choice to proceed with the enrollment or be directed to the free course materials. The purpose of this experiment was to improve the overall student experience and improve coaches’ capacity to support students are likely to complete the course.

Experiment Design

Metric Choice

Invariant Metrics

Number of Cookies

Since this is the number of unique cookies to view the course overview page, we expect this value to be similar in both the experiment and control groups. Since this value will be similar for both cases, it can be considered as an invariant metric.

Number of Clicks

The number of clicks is the unique cookies to click the “Start free trial” button and happens before the free trial screener is shown to the user. Since this is a measure of the number of clicks before the free trial screener, we expect the count to be equal between the experiment and control groups, thus making this an invariant metric.

Click-Through-Probability

This metric is dependent on the number of unique cookies to click the “Start free trial” and the unique cookie to view the course overview page. Both these instances occur before the free trial screener and thus we would expect this metric to be similar between the experiment and control groups.

Evaluation Metrics

Gross Conversion

This metric is the ratio of user-ids to complete checkout and enroll in the free trial and number of unique cookies to click the “Start free trial” button. This is considered an evaluation metric because we expect this conversion rate to be correlated to the overall student experience and improvement of coaches’ capacity. When determining whether or not to launch the experiment, we are looking for a minimum difference of 0.01 between the experiment and control groups.

Retention
This metric is the ratio between the number of user-ids to remain enrolled past the 14-day boundary and the number of user-ids to complete checkout. We would like to see this metric larger in the experiment group as compared to the control group as it indicates overall improved student experiment as well as coaches' capacity to support students. We would like to see a minimum difference of 0.01 between the experiment and control groups.

Net Conversion
This metric is the ratio between the number of user-ids to remain enrolled past the 14-day boundary and the number of unique cookies to click the "Start free trial" button. Again, we would like to see metric this increased for our experiment group when compared to the control group with a minimum difference of 0.0075.

In conclusion, in order for us to launch the experiment, we would like to see an increase in retention, a decrease in gross conversion and the net conversion not to decrease. We would like to see all 3 of these metrics move in the above-mentioned direction in order to launch.

Neither

Number of User-IDs
This is the number of user-ids that enroll in the free trial. This is not considered a valid invariant or evaluation metric. Firstly, it would not be invariant since the user-id is tracked from the point that the student enrolls in the free trial and thus will not be equal between the experiment and control groups. This should not be used as an evaluation metric either since it does not provide any insight into the overall satisfaction of the students.

Measuring Standard Deviation

Gross Conversion Std.	Retention Std.	Net Conversion Std.
0.0202	0.0549	0.0156

I would expect the analytic estimates for gross conversion and net conversion standard deviation values to be accurate. That is because the gross conversion and the net conversion are both the unit of diversion as well as the unit of analysis. When this is the case, the analytic estimates for standard deviation tend to be accurate. For retention however, we would want to collect an empirical estimate of the variability if we had the time.

Sizing

Number of Samples vs. Power

I will not be using the Bonferroni correction.
With an alpha of 0.05 and a beta of 0.2, below are the page views total that we would need to collect to adequately power the experiment.

Gross Conversion

Baseline Conversion Rate	20.625%
Minimum Detectable Effect	1.00%
Beta	20.00%
1 - Beta	80.00%
Alpha	5.00%
Sample Size	25,835
Page Views	645,875

Retention

Baseline Conversion Rate	53.00%
Minimum Detectable Effect	1.00%
Beta	20.00%
1 - Beta	80.00%
Alpha	5.00%
Sample Size	39,115
Page Views	4,741,212

Net Conversion

Baseline Conversion Rate	10.931%
Minimum Detectable Effect	0.75%
Beta	20.00%
1 - Beta	80.00%
Alpha	5.00%
Sample Size	27,413
Page Views	685,325

Therefore, we need total 4,741,212 page views.

Duration vs. Exposure

If we were to divert 100% of the traffic to the experiment, the table below shows the required duration (days) for each evaluation matric:

Gross Conversion	16.1
Retention	118.5
Net Conversion	17.1

For retention, we would have to run the experiment for 119 days, which is too lengthy for us. Given that, we are only left with Gross Conversion and Net Conversion as our evaluation metrics. Regarding the risk assessment of the experiment, it can be safely concluded that no one would get hurt during the duration of the experiment and we are not dealing with any

sensitive data, which limits the risks involved in running the experiment. Since the risk level of this experiment is quite low, it would be reasonable to divert 100% of the traffic and thus would need to run the experiment for 18 days.

Experiment Analysis

Sanity Checks

Below is the table containing the observed values along with the lower and upper 95% Confidence Intervals:

Metric	Expected Value	Observed Value	Lower Bound	Upper Bound	PASS
Number of Cookies	0.500	0.5006	0.4988	0.5012	YES
Number of Clicks on "Start free trial"	0.500	0.5005	0.4959	0.5041	YES
Click-through-probability on "Start free trial"	0.0821	0.0822	0.0812	0.0830	YES

All of the invariant metrics pass the sanity check.

Result Analysis

Effect Size Tests

Metric	dmin	Observed Difference	Lower Bound	Upper Bound	Statistical Significance	Practical Significance
Gross Conversion	0.0100	-0.0206	-0.0291	-0.0120	YES	YES
Net Conversion	0.0075	-0.0049	-0.0116	0.0019	NO	NO

Sign Tests

The p-values for the sign tests are outlined below:

Gross Conversion: 0.0026
Net Conversion: 0.6776

This indicates that gross conversion is statistically significant whereas net conversion is not statistically significant.

Summary

This experiment consisted of diverting cookies that visited the Udacity homepage into one the experiment and control groups. For sanity checking purposes, the invariant metrics that were selected were the Number of Cookies, Number of clicks on “Start free trial” and Click-Through-Probability. The evaluation metrics were Gross Conversion and Net Conversion, where Gross Conversion is the rate of enrollment per cookie and Net Conversion is the rate of payment per cookie. The null hypothesis is that the evaluation metrics are equal in both the experiment and control groups. In order to reject this null hypothesis and accept the alternative hypothesis, the differences in the evaluation metrics for the two groups must exceed the statistically and pre-specified practically significant thresholds. The Bonferroni correction was not used in our methodology since we require all of our evaluation metrics to be significant. The results show

that the invariant metrics pass our sanity check since the metrics show no significant difference between the experiment and control groups at the 95% Confidence Interval. Although the Gross Conversion was found to be statistically significant at the 95% Confidence Interval, the Net Conversion however was **not** found to be statistically significant at the 95% Confidence Interval.

Recommendation

The purpose of this experiment was to analyze whether or not adding the suggestion pop-up would lead to the improvement of overall student experience and better use of resources. Although Gross Conversion was found to be significantly different in the experiment group, this was not the case for Net Conversion. This means that there was a decrease in enrollment but that did not translate to an increase of the number of students that end up making a payment. As a matter of fact, the Confidence Interval of the Net Conversion includes the negative values as well, meaning that Net Conversion may possibly move in the negative direction of what we desire. My recommendation would be to **not** launch but instead pursue further experiments.

Follow-Up Experiment

My suggestion of the follow-up experiment would fall under the same theme of filtering out students that are more likely to cancel their enrollment and potentially waste resources. This filtering process however, needs to be slightly more involved than asking one simple question and offering a suggestion. One major difference in this experiment is that the filtering process would take place after the student has enrolled into a course. Right after enrollment, the student would be diverted to either the experiment or control group, based on the user-id. The experiment group would have to fill out a short quiz that asks several questions in an attempt to gauge the probability that the student would finish end up completing the course and making efficient use of resources such as the coaches' time. There quiz would consist of questions such as "How many hours/week do you work?", "Do you do any volunteering that could impact your time for studies?", etc. After completion, the student would be offered a suggestion whether or not it is appropriate for them to continue and offer advice on how to be successful in completing the course. The **null hypothesis** of this experiment is that the experiment group will not increase retention. The **unit of diversion** is the user-id since the students are split into experiment or control groups after enrollment. The **invariant metrics** would be user-id and the **evaluation metric** would retention. A statistically and practically significant retention rate between the experiment and control group would indicate a launch.