A/B Test for Reducing Early Course Cancellation

Udacity A/B Test Final Project, December 2016

---

1 Experiment Description and Design

Udacity courses currently have two options on the home page: "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.

In 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. 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 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.

The initial unit of diversion to the control and experiment groups is a unique cookie. However, once a student enrols in the free trial, they are tracked by user-id. The same user-id can’t enrol more than once. Users who don’t enrol are not tracked by user-id. Note that the uniqueness of a cookie is determined per day.

1.1 Metric Choice

Invariant Metrics: number of cookies, number of clicks, click-through-probability.

Evaluation Metrics: gross conversion, retention, net conversion.

1.1.1 Invariant Metrics

An invariant metric should not change across experimental and control groups.

- **Number of Cookies**: The number of unique cookies to visit the course overview page. This is the approximation of unique page views and it is the unit of diversion. Because

the pop-up window occurs after clicking on the “start free trial” button, the number of page views, which is the number of cookies here should remain invariant during the experiment.

- **Number of Clicks**: The number of users (tracked as unique cookies at this stage) to click the free trial button. Equal distribution amongst the experiment and control groups would be expected since at this point in the funnel the experience is the same for all users and therefore elements of the experiment would not be expected to impact clicking the “start free trial” button.
- **Click-through-probability**: Unique cookies to click the “start free trial” button per unique cookies to view the course overview page. It should remain unchanged during the experiment, because cookies for page view and clicks happen before the popup window.

1.1.2 Evaluation Metrics

Evaluation metrics are expected to change over the course of the experiment. By comparing differences between the control and experimental groups, we can measure the effect of the screener and test our hypothesis. Each evaluation metric is associated with a minimum difference (d_{min}) that must be observed for consideration in the decision to launch the experiment.

- **Gross Conversion:** This is the number of user-ids to complete checkout and enrol in the free trial per unique cookie to click the “start free trial” button. This metric could measure whether or not less students enrolling in the free trial.
- **Retention:** This is the number of user-ids to remain enrolled past the 14 day trial period, making at least one payment, per number of user-ids to complete checkout. This metric could measure whether or not more students staying beyond the free trial after the 14 day trial period.
- **Net Conversion**: This is the number of user-ids to remain enrolled past the 14 day trial, making at least one payment, per the number of unique cookies to click the “start free trial” button. This metric could measure whether or not the screener had any effect on the 14-day completion rate

In order to launch the experiment, we would expect to reduce frustrating students, without reducing the students to continue past the free trial. With this in mind, in order to consider launching the experiment either of the following must be observed:

- Gross conversion to be increased, while net conversion not to be decreased: the number of enrolled students would be reduced, while the number of students staying beyond the free trial would not be reduced.
- Increased retention: bigger proportion of students staying beyond the free trial in the experiment group.

1.1.3 Other Unused Metrics

- **Number of user-ids**: The number of users to enrol in the free trial. This is for sure not a suitable invariant metric. As for an evaluation metric, it is not ideal. This is because user-id alone is a count, and gross conversion is a fraction that incorporates user-id while also offering a better way to track the effect.

1.2 Measuring Standard Deviation

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

The calculations of the above stand deviation were recorded in the “Data & Calculations” Spreadsheet.

For gross conversion and net conversion, the analytical standard deviation tends to be near the empirically determined standard deviation, because the unit of diversion is equal to the unit of analysis. However, for retention, this is not the case. If we do ultimately decide to use retention as an evaluation metric, we’d better calculate the empirical variability.

1.3 Sizing

1.3.1 Number of Samples vs. Power

My initial approach will not deploy the Bonferroni correction, that decision will be made based on my final choice of evaluation metrics and associated criteria. To know the exact number of pageviews required for the experiment, we need to calculate the sample size that we will need for each evaluation metric. The calculation was recorded in the “Data & Calculations” spreadsheet.

	Gross Conversion	Retention	Net Conversion
Baseline Conversion	20.63%	53%	10.93%
Minimum Detectable Effect	0.01	0.01	0.0075
Alpha	0.05	0.05	0.05
Beta	0.2	0.2	0.2
Sample Size	25,835	39,115	27,413

	Gross Conversion	Retention	Net Conversion
Pageviews	645,875	4,741,212	685,325

The number of page views needed to power the experiment appropriately is 4,741,212.

1.3.2 Duration vs. Exposure

Given 70% diversion, we would need around 25 days to run the experiment.

If we divert 100% of the traffic, given 40,000 pageviews per day, the experiment would take 119 days. This is too long for an experiment and we should reduce the duration. A 119 day experiment with 100% diversion of traffic presents both a business risk (potential for: frustrated students, lower conversion and retention, and inefficient use of coaching resources) and an opportunity risk (performing other experiments).

We can exclude retention as an evaluation metric and consider the next limiting metric, net conversion. This reduces the number of the required pageviews to 685,325, and it would take18 days given 100% diversion assuming there were no other experiments running simultaneously. In general, this is not a risky experiment as the change would not be expected to cause a precipitous drop in enrolment, it does not affect existing paying customers, and it is simple enough that there is a low chance of bugs occurring in the process.

Nevertheless, 100% diversion may be scaled down depending on other experiments of interest to be performed concurrently. Therefore, I would divert 70% of the traffic to the experiment, and the experiment will take 25 day given that diversion rate.

2 Experiment Analysis

2.1 Sanity Checks

The invariant metrics were tested at the 95% confidence interval. The calculations were recorded in the “Data & Calculations” spreadsheet.

Invariant Metrics	Value in Control	Value in Experiment	Total	Expected Value (Central of CI)	CL Lower	CL Upper	Observed Value	Result
Number of Cookies	345543	344660	690203	0.5	0.4988	0.5012	0.5006	Pass
Number of Clicks	28378	28325	56703	0.5	0.4959	0.5041	0.5005	Pass
Click-through-probability	0.0821	0.0822		0.0821	0.0812	0.0830	0.0822	Pass

All of the invariant metrics passed the sanity check.

2.2 Result Analysis

2.2.1 Effect Size Tests

For each evaluation metric, both statistical and practical significance were tested. The minimum detectable effect is the smallest difference that we will accept between experimental and control groups in order to be practically significant. The metrics were tested at the 95% confidence interval for the difference between the experiment and the control groups. The calculations were recorded in the “Data & Calculations” spreadsheet.

Evaluation Metrics	Value in Control	Value in Experiment	Total	Difference (Central of CI)	CL Lower	CL Upper	Dmin	Statistical significance?	Practical Significance?
Gross Conversion	0.2189	0.1983	0.2086	-0.0206	-0.0291	-0.0120	0.01	YES	YES
Net Conversion	0.1176	0.1127	0.1151	-0.0049	-0.0116	0.0019	0.0075	NO	NO

Gross conversion is both statistically and practically significant. Net conversion is not statistically significant or practically significant.

2.2.2 Sign Tests

To further test each of our evaluation metrics, I conducted a binomial sign test using the day by day data. Again, the calculations for the sign tests were recorded in the “Data & Calculations” spreadsheet.

Evaluation Metrics	Number of Experiments	Number of Successes	Alpha	P value	Statistical Significance?
Gross Conversion	23	4	0.05	0.0026	YES
Net Conversion	23	10	0.05	0.6776	NO

2.2.3 Summary

The effect size tests determine that gross conversion is both statistically and practically significant, while net conversion is neither. The requirement for launching the experiment is that the null hypothesis must be rejected for all evaluation metrics and that the difference between branches must meet or exceed the practical significance threshold.

The Bonferonni correction is not appropriate because our acceptance criteria requires statically significant differences for all evaluation metrics. The Bonferonni correction is a method for controlling for type I errors (false positives) when using multiple metrics in which relevance of any of the metrics matches the hypothesis. In this case the risk of type I errors increases as the number of metrics increases (significance by random chance). In our case in which all metrics must be relevant to launch, the risk of type II errors (false negatives) increases as the number of metrics increases, so it stands to reason that controlling for false positives is not consistent with our acceptance criteria.

The sign tests allow for an additional form of analysis. The conclusion from the sign test mirrors that of the effect size test, that gross conversion is significant but net conversion is not. In this case, both tests agree and our conclusions with regard to both metrics are strongly supported.

2.3 Recommendation

Gross conversion turned out to be negative and practically significant. The screener was proved to be effective at reducing the number of people to continue from click to enrol, which means, reducing the number of frustrating students. This is a good outcome because we lower our costs by discouraging trial signups that are unlikely to convert.

Unfortunately, Net conversion ended up being statistically and practically insignificant. Moreover, the confidence interval includes negative numbers, which means there is a risk that the screener may lead to a decrease in net conversion, which is to say, it might reduce the number of students to continue past the free trial and eventually complete the course.

Considering this, my recommendation is not to launch, but rather to pursue other experiments.

3 Follow-Up Experiment

In the current experiment, if a student indicates less than the recommended number of hours, 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. This experiment was proved to effectively decrease gross conversion, which means, it would reduce frustrating students who enrol. But there is a risk that it might reduce the number of students to continue past the free trial and eventually complete the course.

In order to increase the number of students who stay beyond the free trial, I would suggest provide bigger motivation for the students who enrolled for the free trial and encourage them to stay beyond the free trial and eventually complete the course.

The follow-up experiment can add encouraging messages after the students enrol for the free trial, such as the advantages after completing the whole course, stories or videos about people who have completed this same course getting related job offers, etc,. In this

way, students who enrolled in the free trial would be expected to have bigger motivation to past the free trial and eventually complete the course.

The **hypothesis** is that this kind of encouraging messages would motivate some students who might otherwise drop out during the 14-day trial to continue past and possibly complete the course. It would also not affect those people that would otherwise continue through the trial and complete the course had there been no pop-up message.

Considering this design, **retention** would be the most suitable **evaluation metric** to test our hypothesis. We would divert traffic evenly among the control and experimental groups, and the **unit of diversion** would be the **user-ids**. The most suitable **invariant metric** would be the **number of user-ids** to complete checkout and enrol in the free trials.

If retention is statistically and practically significant at the end of the experiment, we can laugh the new feature.