A/B Testing on Udacity's Free Trial Screener: Design and Analysis

Experiment Design

Metric Choice

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

***Evaluation Metrics**: gross conversion, retention, net conversion*

Invariant Metrics

Invariant metrics are metrics that we expect to not have any significant changes between the control and the experiment. They remain unaffected by the introduction of the free-trial screener Udacity is testing. I expect similar distributions for the following metrics in both the control and experiment groups.

Number of Cookies: The number of unique cookies to view the course overview page. This is the basis of each individual subject in our experiment before the users enroll in the free trial. Hence, this is also our unit of diversion. I expect even distribution of cookies between the control and experiment groups without being influenced by the screener. It is therefore appropriate as an invariant metric.

Number of Clicks: The number of unique cookies to click the "Start free trial" button. This happens before the free trial screener is triggered which is why we expect equal distribution amongst the experiment and control groups.

Click-through-probability: The number of unique cookies to click the "start free trial" button divided by number of unique cookies to view the course overview page. At this point in the funnel, the user experience is same for both the control and experiment groups. That is why Click-through-probability(CTP) is a good choice for an invariant metric.

Evaluation Metrics

Evaluation metrics are metrics we expect to be influenced by the introduction of the Free trial screener in the experiment group . There is a possibility that our evaluation metrics will have different distributions between our 2 groups. Each evaluation metric has a minimum difference (dmin) that must be exceeded for the change to be practically significant.

Gross Conversion: This is the number of user-ids to complete checkout and enroll in the free trial divided by the number of unique cookies to click the "start free trial" button.(The practical minimum difference (dmin) = 0.01 or 1%)

Retention: This is the number of user-ids to remain enrolled past the 14 day trial period (and thus, make at least one payment) divided by the number of user-ids to complete checkout. (The practical minimum difference (dmin) = 0.01 or 1%)

Net Conversion: The number of user-ids to remain enrolled past the 14 day boundary (and, thus make at least 1 payment) divided by the number of unique cookies to click the "Start free trial" button. (The practical minimum difference (dmin) = 0.0075 or 0.75%)

The objective of the Udacity experiment is to minimize student frustration by making them aware of the student expectations beforehand when it comes to time commitment. This would allow Udacity to improve the coaches' capacity to support the students who are more likely to stay enrolled past the free trial period and complete the course, consequently improving the overall the student experience as well.

Unused Metrics: *Number of user-ids*

Number of user-ids: The number of users who enroll in the free trial. User-ids are tracked only after enrolling in the free trial, and the number of users who enroll in the free trial is dependent on the experiment itself and therefore, equal distribution between the control and experimental groups is not expected. That is why it is not an appropriate choice for an invariant metric. The number of visitors who click the start-free trial can fluctuate on a day-to-day basis, which can influence the distribution of user-ids in the control and experiment groups. This can eventually skew our results of the effect of the checkout page rendering. User-id count could be used to evaluate how many enrollments stayed beyond the 14 day free trial boundary, but since it isn't normalized, I have elected not to use it. Other metrics, like the gross conversion gives me similar information as this metric while marginalizing the effect of the number of "start free-trial" clicks.

We expect the gross conversion will decrease significantly, which indicates that the operational cost of maintaining resources and student support services will be lower by introducing the new screener; while net conversion will not decrease significantly, which indicates that applying the screener will not reduce Udacity's revenues. If both of these hypotheses hold true, we can launch the experiment.

Measuring Standard Deviation

Analytical Estimate of Standard Deviation

Evaluation Metric	Standard Deviation
Gross Conversion	.0202
Retention	.0549
Net Conversion	.0156

The analytical estimate of standard deviation tends to be almost equal to the empirically calculated standard deviation for cases where the unit of diversion is the same as the unit of analysis. For Gross Conversion and Net Conversion, both of these units are the “Number of cookies”. For Retention though, it is not so we should calculate the empirical variability of this metric if we decide to keep it as one of the evaluation metrics.

Sizing

Number of Samples vs. Power

I did not deploy the Bonferroni correction because the evaluation metrics in the test are considerably correlated and the Bonferroni correction will be too conservative. Pageviews required for each metric were calculated using an alpha value of 0.05 and beta value of 0.2.

Pageviews for Each Evaluation Metric to Achieve Target Statistical Power

Gross Conversion

- Baseline Conversion: 20.625%
- Minimum Detectable Effect: 1%
- alpha: 5%
- beta: 20%
- Statistical Power(1 - beta): 80%
- sample size = 25,835 enrollments/group
- Number of groups = 2 (experiment and control)
- total sample size = 51,670 enrollments
- clicks/pageview: 3200/40000 = .08 clicks/pageview

- pageviews = 645,875

Retention

- Baseline Conversion: 53%
- Minimum Detectable Effect: 1%
- alpha: 5%
- beta: 20%
- Statistical Power(1 - beta): 80%
- sample size = 39,155 enrollments/group
- Number of groups = 2 (experiment and control)
- total sample size = 78,230 enrollments
- enrollments/pageview: 660/40000 = .0165 enrollments/pageview
- pageviews = 78,230/.0165 = 4,741,212

Net Conversion

- Baseline Conversion: 10.9313%
- Minimum Detectable Effect: .75%
- alpha: 5%
- beta: 20%
- Statistical Power(1 - beta): 80%
- sample size = 27,413 enrollments/group
- Number of groups = 2 (experiment and control)
- total sample size = 54,826
- clicks/pageview: 3200/40000 = .08 clicks/pageview
- pageviews = 685,325

Pageviews Required: 4,741,212

Duration vs. Exposure

Based on our calculation above, even if we divert 100% of our traffic to the experiment, the high number of pageviews required makes this experiment at least 119 days long which is too long for an experiment like this which possibly affects user experience and retention, conversion rates. Hence, I decide to remove Retention as an evaluation metric. This reduces our required number of pageviews to 685,325.

Updated Evaluation Metrics*: gross conversion, net conversion*

Updated Pageviews required*: 685,325*

I decide to redirect 60% of the traffic to our experiment, and the length of the experiment is therefore `685325 / (40000 * 0.6) = 28.6` days (where 40000 is the

baseline number of visitors per day). We approximately use 29 days as the length of the experiment.

The 60% traffic being redirected to the experiment means that 30% will go to the control group while the other 30% is assigned to the experiment group, and therefore we risk about a third of the users exposed to a different enrollment experience. We divert a fairly high percentage of all traffic to the experiment to keep the length of our experiment short(less than 30 days). The good thing about the experiment though is that it doesn't affect the already enrolled students or the students who are highly motivated and expect to commit time. The experiment doesn't collect any personal sensitive information which is why it is not a high risk experiment. The only possible risk from the experiment is financial however, that is one of the things we are directly monitoring for this experiment, so it's not as much of a concern because we will see decline pretty quickly. Plus, the duration of this experiment is less than a month so the financial risk is not high. Therefore, overall the whole experiment would not considered highly risky.

Experiment Analysis

Sanity Checks

For the invariant count metrics, we expect equal dispersion of cookies and clicks into the experiment and control groups as they model a Bernoulli distribution with probability 0.5. For the CTP metric, we use the CTP of our control group as the expected value for our sanity check. We will test this at the 95% confidence interval.

Metric	Expected Value	Observed Value	CI Lower Bound	CI Upper Bound	Result
Number of Cookies	0.5000	0.5006	0.4988	0.5012	Pass
Number of clicks on "start free trial"	0.5000	0.5005	0.4959	0.5042	Pass
Click-through-probability	0.0821	0.0822	0.0812	0.0830	Pass

Result Analysis

Effect Size Tests

I will compute the 95% Confidence interval for the difference observed between the experiment and control groups for each of the evaluation metrics. A metric is considered statistically significant if the confidence interval does not include 0. It is considered practically significant if the interval doesn't include the practical minimum difference.

Metric	dmin	Observed Difference	CI Lower Bound	CI Upper Bound	Result
Gross Conversion	0.01	-0.0205	-.0291	-.0120	Statistically and Practically Significant
Net Conversion	0.0075	-0.0048	-0.0116	0.0019	Neither Statistically nor Practically Significant

Sign Tests

For each evaluation metric, we use a day-by-day breakdown as the basis for our sign test. We use the online calculator to perform the sign test and calculate the p-value for each metric.

Metric	p-value for sign test	Statistically Significant (alpha .05)?
Gross Conversion	0.0026	Yes
Net Conversion	0.6776	No

Summary

As the number of metrics increase, it becomes increasingly likely that one of the metrics will be a false positive. This is where Benferroni correction comes in handy and prevents us making the wrong launch recommendation. However, in this experiment, we would only launch if all evaluation metrics show a significant change which allows us to not use Benferroni correction in this case. The effective size and sign test results align with each other and show us that the decrease in gross conversion is practically significant while the change in net conversion is not significant at all.

Recommendation

This experiment was designed to reduce the number of frustrated students who enrolled and left during the free trial, improve overall student experience by increasing the coaches' capacity to support students who are more likely to stay enrolled and finish the course, and finally, have a higher proportion of students who eventually pay. In order to achieve these, Udacity tries incorporating a "free-trial" screener that cautions users about the expected time commitment. From the tests above , we can see that there is a statistically and practically significant decrease in Gross Conversion but no significant differences were observed in Net Conversion. In fact, the confidence interval of the net conversion includes negative numbers, which makes it possible that this number goes down by an amount that can actually hurt the business. This means there was a decrease in enrollment considerable enough to reduce costs from the business standpoint, but students were no more likely than before to stay enrolled and convert to paying customers. On the contrary, there is a chance that net conversion can go down just enough to decrease Udacity's revenue.
I believe this could be due to the fact that the trial is free after all and even if students are unsure of wanting to complete the course, they can always enroll and find out for themselves as to whether they are interested in the course material or how much time it personally takes them to progress through the course. The filtering did reduce the number of students who signed up for the free trial and I suspect most of these are students who already knew that they didn't want to pay for the course and/or were planning on committing a lot less than 5 hours a week for the coursework. Overall, the experiment failed to achieve the main objectives and as such, my recommendation would be to not launch and instead explore other hypotheses and experiments.

Follow-Up Experiment

The original experiment assumed that student frustration led to them cancelling the enrollment but there could be lot of other factors why students decide to cancel. For example, they might not have liked the course content or it wasn't what they were

looking for, they don't have prerequisite knowledge for the course, they don't think it is worth the money or they might not have that money to spend at all.

As part of the new experiment group, I would recommend making and adding a short 2 week intro-course within the course work that goes over some of the advantages of taking the paid course option over accessing the course material for free and why it would benefit the students, going over the prerequisite topics needed to better understand the future coursework later while giving high-level real life application based introductions to each main lesson/project/course. This could also end up increasing engagement time for a student in the first 2 weeks which we can test but not consider as a factor for launching. We will also include the time screener on the "Free-trial" checkout page for both the control and experiment groups, thereby slightly reducing the number of students who don't intend to or aren't able to complete the course beforehand.

My hypothesis is that incorporating this new intro-course will not only increase engagement time but also provide a substantial foundation for the upcoming lessons, which in turn should reduce student frustration and help increase the retention rates.

Setup: Students will either be randomly assigned(after enrollment) to a control group in which they are funnelled into the original coursework, or an experiment group in which they are directed to the coursework that now has a mandatory 2 week intro.

Unit of Diversion: The unit of diversion will be user-id as the change takes place after the enrollment. Until the enrollment, the users are tracked by cookies. The same user-id cannot enroll in the "free-trial" twice and for users who do not enroll, their user-ids are not tracked.

Invariant Metrics: The invariant metric will be user-id, since I expect an equal distribution between the experiment and control groups.

Evaluation Metrics: The evaluation metric will be Retention because we are interested in retaining more students who tried the coursework during the "free-trial" period and this change in retention can be attributed to the mandatory intro-course in the experiment group. A statistically and practically significant increase in Retention would mean that our experiment is successful in achieving the objectives set earlier.

If the rate of Retention has increased and the difference is practically significant at the end of our experiment, we can launch and roll out the newly updated coursework for all the students interested in that course. If we plan on implementing this design for all the programs, we should ramp up gradually from program to program.

References:

- Introduction to A/B Testing (Udacity)
- GraphPad