A/B Testing

Experiment Overview

Udacity courses currently have two options on the home page: “start free trial” and “access course materials”. If the students click ‘start free trial’, they will be asked to enter their credit card information, and then they will enroll in a free trial for a paid version. After 14 days, they will automatically be charged unless the students canceled first. If the students click on “access course materials”, they will be able to view the videos and take the interactive quizzes for free, but they will not receive a verified certificate or one-on-one coaching support, moreover, they will not be able to submit their final projects and get feedbacks for them.

In this experiment, Udacity tested a change where if the student clicked “start free trial”, they were asked how much time they had available to commit to the course. If the students have 5 or more hours per week, they would go through the registration process as usual. However, if not; a message would appear indicating that Udacity course usually require a greater time commitment for successful completion. Students will also be told that they can access the course materials for free.

The hypothesis for this experiment was that this might set a clear expectation for students upfront, thus reducing the number of frustrated students who left the free trial because they did not have enough time for the course- without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this works, Udacity could improve the overall student experience and improve the counseling’s capacity to support students who are likely to complete the course.

Experiment Design

Metric Choice

Invariant Metric: Number of cookies, Number of clicks, Click-through-probability
Valuation Metric: Gross Conversion, Net Conversion, Retention

- Number of cookies: Good invariant metric because number of cookies is not going to be affected by the change that company is launching at the time of enrollment.
- Number of user-ids: Not a good invariant nor evaluation metric. Since the enrollment might depend on the ‘start free trial' page, we could expect to see different values in control and experiment group. Therefore, it can’t be invariant. It is not a good evaluation metric because it is reductant to other metrics such as gross conversion. Gross conversion is a fraction of user-id and using gross conversion is a better choice.

- Number of clicks: Similar to number of cookies. Good invariant metric because the clicks happen before the user sees the page before they decide to click on the button.
- Click-through-probability: invariant metric. Again, since the users have not seen the page we tested on before they decide to click the button, the click-through-probability also does not depend on our test and is a good invariant metric.
- Gross conversion: Not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. Good evaluation metric because it is directly dependent on the effect of the experiment and allows us to show whether we managed to decrease the cost of enrollments that aren't likely to become paying customers.
- Retention: Not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. Good evaluation metric because it is directly dependent on the effect of the experiment, and shows positive financial outcome of the change.
- Net conversion: Not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. Good evaluation metric because it is directly dependent on the effect of the experiment, and shows positive financial outcome of the change.

I will look at both Gross Conversion and Net Conversion. The gross conversion will show us whether we lower our cost by introducing new pop up. Net conversion will show how the change affects our revenue. After the experiment, we should expect that gross conversion have a significant decrease, and net conversion should not decrease significantly.

Measuring Standard Deviation

To determine whether the analytical estimates of standard deviation are accurate, such as whether it matches the empirical standard deviation, we consider whether or not the unit of analysis and unit of diversion matches up.

Using the online calculator, we calculated number of samples required as following:

For 5000-page view:

- number of clicks = $5000 \times 0.08 = 400$
- number of enrollment = $5000 \times 0.08 \times 0.20625 = 82.5$

Baseline table:

Unique cookies to view page per day	40000

Unique cookies to click “Start free trial” per day	3200
Enrollments per day	660
Click-through-probability on “Start free trial”	0.08
Probability of enrolling, given click	0.20625
Probability of payment, given enroll	0.53
Probability of payment, given click	0.1093125

Metric	Value	Std	Std /5000	Probability
Gross conversion	0.2063	0.0072	0.0202	0.0800
Retention	0.5300	0.0194	0.0549	0.0165
Net conversion	0.1093	0.0055	0.0156	0.0800

- Gross Conversion: The unit of diversion = unit of analysis. Analytical estimate of Standard Deviation tends to be empirical estimate of Standard Deviation.
- Retention: Retention unit of diversion is not the same as unit of analysis.
- Net Conversation: The unit of analysis and unit of diversion are both the same for "gross conversion" metric, and the analysis directly applies here. The analytical estimate is expected to be mostly accurate, but collecting more data to verify if one has time will be even better.

Sizing

Number of Samples vs. Power

I want gross conversion significantly decrease AND net conversion does not significantly decrease.
I will not use Bonferroni as it is too conservative. I want all my metrics to be significant.
$\alpha = 5\%$, $\beta = 20\%$

- Gross Conversion (base conversion rate= 20.625%, dmin=1%)
- Retention (base conversion rate=53%, dmin=1%)

- Net conversion (base conversion rate=10.93125%, dmin=0.75%)

From this calculator, we get samples show below.

Metric	Pageviews	Sample size(clicks)
Gross conversion	322,937	25835
Retention	2,370,606	39115
Net conversion	342,662	27413

We need pageviews:

- Gross Conversion: 25835 x 40000 / 3200 =322,937
- Retention: 39115 x 40000 / 660 = 2,370,606
- Net conversion: 27413 x 40000 / 3200 = 342,662

Duration vs. Exposure

First trial (Use Gross Conversion, Retention, and Net Conversion as evaluation metrics):

- Number of page views = 2370606 x 2 = 4741212 (because two groups)
- Fraction = 1.0
- Days = 4741212 / 1 / 40000 = 118

First trial requires 118 days to do. This is way too long.

Second trial (only Gross Conversion and Net Conversion as evaluation metrics):

- Number of page views = 342662 x 2 = 685324 (because two groups)
- Fraction = 1.0
- Days =685324 / 1/40000 = 18

Second trial requires 18 days to do.

Experiment of 18 days is short enough, so we would choose Gross Conversion and Net Conversion as evaluation metrics.

In this experiment, when students want to enroll, we ask how much time they are willing to commit for the course they want to enroll and recommend students not to enroll if they could not investigate more than 5 hours time. This option works for student as those who don't have enough time could choose to

access course materials, do quizzes, or watch videos whenever. In addition, when they decide to enroll, they could do it at any time. We also do not ask personal information in this experiment so privacy is not an issue. This experiment does not affect the database nor the design of the website so the website is not harmed by the experiment, either. Fraction of 1.0 is chosen because if we decrease the fraction of traffic, we would need more time to run this experiment, and 17 days isn't particularly short either.

Experiment Analysis

Sanity Checks

	Control Group	Experiment
#pageview	345543	344660
#clicks	28378	28325

Metric	Lower Bound	Upper Bound	Observed Value	Result
# of Cookies	0.498820392149	0.501179607851	0.5006396669	PASS
# of Clicks	0.495884957	0.5041155043	0.5004673473	PASS
Click through Probability	0.08121035975	0.0830412674	0.08212581357	PASS

Result Analysis

Effect Size Tests

95% Confidence interval around the difference between the experiment and control group for evaluation metrics.

Metric	dmin	Lower Bound	Upper Bound	Statistical Significant	Practical Significant
Gross Conversion	1%	-0.0291	-0.0120	TRUE	TRUE
Net Conversion	1%	-0.0116	0.00185	FALSE	FALSE

Gross Conversion is both Statistically Significant and Practical Significant
Net Conversion is neither Statistically Significant and Practical Significant

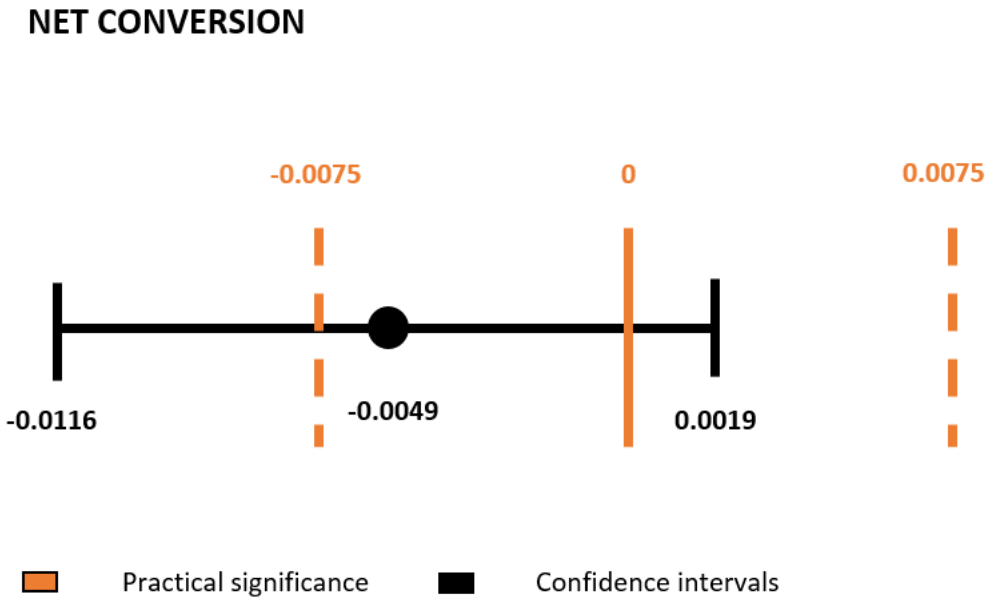
Sign Tests

Metric	p-value	Statistically Significant
Gross Conversion	0.0026	Yes
Net Conversion	0.6776	No

Summary

Bonferroni correction is not used here because our launch decision is based upon two metrics, Gross Conversion and Net Conversion. To launch, we need both the metrics to meet the expectations; that is, we want gross conversion significantly decrease AND net conversion does not significantly decrease. Bonferroni correction is suitable when it is applied to 'OR' situation. There were no discrepancies between the effect size hypothesis tests and the sign tests.

Recommendation



Based on the analysis, gross conversion turned out to be negative and practically significant. This is good because this decrease the costs by discouraging trial signups that are unlikely to convert. Net conversion, however, ended up being statistically and practically insignificant. Refer to the illustration above, the confidence interval includes the negative practical significance boundary. That is, it's possible that this number went down by an amount that would matter to the business. This means that there is a risk that

the introduction of the trial screener may lead to a decrease in revenue. This is not an acceptable risk to launch.

This experiment makes me consider testing other designs of the screener before we decide whether to release the feature.

Follow-Up Experiment

Goal of a company is to earn money by satisfying customers. In this experiment, we tried to filter out the users who are going to enroll to trial but are not going to spend a lot of time on studying.

In follow-up experiment, we can perform another experiment by changing the number of hours from screener to prerequisite knowledge required to start the course. This screen aims to help students to get an idea about what knowledge they should have before joining Udacity course. If the students don't have the knowledge, then the screen should suggest them to go to the suggested courses that are listed in prerequisites and can join those courses.

Null Hypothesis: No significant difference between control and experiment group.
Unit of diversion: Cookie

For testing this hypothesis, we have to measure number of cookies, number of clicks, number of enrollments, and number of payments. From those, we can calculate Gross Conversion & Net Conversion.

If Gross Conversion & Net Conversion will result to be statistically and practically significant then we will be able to launch our test.

References:

- https://rpubs.com/superseer/ab_testing
- Udacity A/B Testing
- Udacity Discussion Forum