P7 - A/B Testing Final Project

Udacity is considering a change in their user interface. New users who attempt to enroll in a free trial will be prompted to answer a question about how much time they are willing to spend on their coursework. Users who enter a low number of hours available are encouraged by the system to access the course materials for free without enrolling.

This change is to “set clearer expectations for students up front, thus reducing the number of frustrated students who left the free trial because they didn’t have enough time”. The change should also avoid “significantly reducing the number of students to continue past the free trial and complete the course”.

Unfortunately we do not have enough time (or data!) to measure those who complete their courses. Instead, we will measure those who continue to use Udacity past their free trial period and make an initial payment for the service.

The null hypothesis is that enrollment (beginning the free trial) and retention (making the first payment) will not be affected by this change.

The alternative hypothesis is that the change will affect either the enrollment, the retention, or both. Let’s run this experiment and evaluate the results.

Experiment Design

Metric Choice

Invariant Metrics

I chose three invariant metrics to monitor in this experiment:

Number of cookies: This metric refers to the number of unique cookies to visit the Udacity Course Overview page each day. Because this metric is generated before a user begins their free trial, it will not be affected by the proposed change. The metric should stay the same for both control and experiment groups. This will be an invariant metric.

Number of clicks: This metric counts unique cookies where the “Start Free Trial” button is clicked. Again, because the events leading to this button are unchanged in the experiment, it should not be different for the control and experiment groups. Therefore this metric is also an invariant metric.

Click-through-probability: This metric is directly proportional to the two metrics above (Number of clicks / Number of cookies). If the number of cookies does not change significantly, and the number of clicks also does not change significantly, then the click-through probability is also not likely to change significantly between the control and experiment groups as the events measured happen before a user sees changes that are part of this experiment. This is also an invariant metric.

Evaluation Metrics

I chose three evaluation metrics for this experiment.

Gross Conversion - This metric measures the proportion of users who enroll in the course with their credit card details. The experiment introduces an additional step to this process where a user is warned against enrolling if they cannot spend significant time on coursework. This may affect the number of users who choose to continue and enroll in the course (as opposed to viewing freely available course videos). Since we expect this number to change in our experiment group, we call Gross Conversion an evaluation metric.

H0 (null hypothesis) = Gross Conversion rate will not change in the experiment group.
HA (alternative hypothesis) = Gross Conversion rate will change in the experiment group.

Retention - This metric measures the proportion of users that enrolled in the free trial who continue to use the service and make at least one payment. This experiment introduces a change in the enrollment process. Since the change affects students enrolling in the Udacity program, the retention of these students may also change. Therefore Retention is another evaluation metric.

H0 (null hypothesis) = Retention rate will not change in the experiment group.
HA (alternative hypothesis) = Retention rate will change in the experiment group.

Net Conversion - This metric measures the proportion of unique cookies representing users who enroll and continue to use the course (making an initial payment). Similar to our retention metric, we expect that changes in the enrollment process may affect those who enroll. Changes to enrollments will affect this metric so we also consider Net Conversion an evaluation metric. This metric is perhaps the most important metric as it represents the overall change in paid users across the control and experiment groups.

H0 (null hypothesis) = Net Conversion rate will not change in the experiment group.
HA (alternative hypothesis) = Net Conversion will change in the experiment group.

Unused Metrics

I decided not to use the **Number of user-ids** metric. User IDs are generated in cookies when enrolling in the free trial.

Since enrollments are likely to change in this experiment, this metric will also change. It is not an invariant metric.

Compared to our other evaluation metrics, when evaluating changes in the experiment group, we are most concerned with enrollments and paying customers. The total numbers of User IDs in the control and experiment groups do not tell us anything about how many free trial-enrolled users have made their first payment. Therefore it is of limited use to us in evaluating the success of our experiment.

Experiment Changes to Metrics

In this experiment, the invariant metrics should stay the same for both control and experiment groups. If there are changes in any of the three invariant metrics, then we lose confidence that the control and experiment groups represent similar samples.

The goal of this experiment is to reduce the number of students who sign up for a free trial then leave the program before making their first payment. This goal can be broken up into two parts:

1) Reduce number of students who sign up for a free trial. These reductions should affect users who would have left the program before their first payment.
2) Maintain the number of students who enroll and pay for a Udacity course.

The experiment goal will be successful if the alternative hypothesis for Retention is correct and the experiment group has a higher retention rate than the control group. We can also check the Gross and Net conversion rates to gauge success in the project. If the gross conversion rate decreases without a similar decrease in net conversion rate, then we can call this experiment a success.

Measuring Standard Deviation

Gross Conversion - The analytic standard deviation is: **0.0202**. Gross Conversion measures the proportion of user IDs to cookies. A single user ID may be represented by more than one cookie. Since the unit of analysis (User IDs) is different from the unit of diversion (cookies), the empirical variability will likely be higher than the analytic variability.

Retention - The analytic standard deviation is: **0.0549**. Retention measures the ratio of users IDs which remain enrolled through first payment. Since the unit of analysis and unit of diversion are both User ID counts, the analytic and empirical variability should be very similar.

Net Conversion - The analytic standard deviation is: **0.0156**. This metric is similar to Gross Conversion. The unit of analysis is total number of user ids. The unit of diversion is cookie count. Since the units are different, the empirical variability will be higher.

Sizing

To estimate the number of pageviews (and cookies required to generate these pageviews), I used a sizing calculator available at this URL:

http://www.evanmiller.org/ab-testing/sample-size.html

Number of Samples vs. Power

Since the three evaluation metrics are all related to each other, a Bonferroni correction is not needed. Gross Conversion (enrollments) affects Retention and Net Conversion. I used an alpha of 0.05 and beta of 0.20 to determine sizing. Here are the results for the required number of page views, cookies required to generate these pageviews, and a sample size large enough for both control and experiment groups.

	# Page Views	# Cookies	Control and Experiment Size	# Days
Gross Conversion	25,835	322,937.50	645,875	17
Retention	39,115	2,370,606.06	4,741,212.12	119
Net Conversion	27,413	342,662.50	**685,325**	18

Duration vs. Exposure

It will take 4.7million cookies to measure the effects of the enrollment changes for Retention. Even if 100% of Udacity's traffic (40,000 cookies per day) is used for the experiment, it will take 119 days (4-months) to get enough data to verify the results of this experiment on Retention (% of enrollments who transitioned to paid customers). This is too long for Udacity and not practical.

Therefore during the sizing process, I decided to remove Retention from the target metrics and focus on Gross Conversion and Net Conversion instead. Now **685,325** unique cookies are required. With **100%** of Udacity's website traffic devoted to this experiment, it will take **18** days to gather enough data for our experiment and control groups. This is a reasonable time frame and will give us both a gross (# of users enrolling in the free trial) and net (# of users enrolling in the free trial and paying) conversion rate.

This experiment is not risky for Udacity or its participants. Users in the experiment group are introduced to a new screen which asks them about their time availability. For those participants who have limited time available, they are encouraged to access course materials free online. These participants can continue to enroll if they wish. A participant's physical, psychological, and economic well-being is not affected by the changes proposed in this experiment.

Experiment Analysis

Sanity Checks

When comparing control and experiment samples, it is important to do a sanity check and see that invariant metrics have not changed between the two samples.

I assume a binomial distribution for the first two invariant metrics (PageViews/Cookies and "Start a Free Trial" Clicks) based on the following assumptions:

1) There are two types of outcomes for these metrics. Cookies are assigned to either the control group or experiment group. For each cookie, users may or may not click on the "Start Free Trial" button. We should see similar counts in both control and experiment groups.
2) The events are independent. Cookies are randomly assigned to each group. Since users are not aware of which group they are in, they are neither more nor less likely to click on'Start Free Trial' after being assigned to a group.
3) There is an identical distribution of values for both experiment and control groups.

Assuming a binomial distribution I calculated the standard deviation of the binomial using a 50% probability and 95% Confidence Interval. Here are my results:

	SD	SE	Mean	Lower Bound	Upper Bound	Mean (Control)
PageViews	0.0006	0.0012	0.5	**0.4988**	**0.5012**	**0.5006**
Clicks	0.0021	0.0041	0.5	**0.4959**	**0.5041**	**0.5005**

The ratio of PageViews and Clicks in the control group is 0.5006 and 0.5005 respectively. This falls between our lower and upper bounds. It passes our sanity check.

The click-through rate on the other hand is not a metric with two kinds of outcomes. Therefore we need to calculate a pooled standard deviation. For this metric, we use the difference in probability (0.0001) to set our lower and upper bounds. If 0 is within these bounds, then we can attribute any differences to statistical chance.

	SD	SE	Mean	Lower Bound	Upper Bound	Click Rate Diff
Click Rate	0.0007	0.0013	0	**-0.0012**	**0.0014**	**0.0001**

The difference in click-through probability is between the lower and upper bounds. Our sanity check passes for this metric as well.

Result Analysis

Effective Size Tests

	SD	SE	d	Lower Bound	Upper Bound
Gross Conversion	0.0044	0.0086	-0.0206	**-0.0291**	**-0.0120**
Net Conversion	0.0034	0.0067	-0.0049	**-0.0116**	**0.0019**

The Gross Conversion rate decreased by 0.0206 (2.06%). That means that there are 2% fewer users completing their free trial enrollment after clicking the “Start Free Trial” button. The 95%CI standard error is 0.0086. Our results state that this decrease exceeds our standard error therefore the reduction is statistically significant. The d(min) value for this metric in the instructions was 0.01. Since the change is down by 2%, we can say that this change is also practically significant.

The Net Conversion also decreased. However this decrease was much smaller (0.49%) and within our margin of error. This change may well have been due to statistical chance. Therefore the change is not statistically significant. The d(min) score for this metric was 0.0075. Since our confirmed difference is also smaller than the d(min) score, this change is **not** practically significant.

Sign Tests

To verify the effects of the experiment, I also ran a sign test on the daily results using a sign calculator available at this URL: http://graphpad.com/quickcalcs/binomial1.cfm

Here are my results:

	# successes	# trials	Two Tailed Result
Gross Conversion	4	23	**0.0026**
Net Conversion	10	23	**0.6776**

For the Gross Conversion data, I noticed that the experiment rates were higher on just 4 of the 23 days. Given that the probability of data being higher or lower is 0.5 (flip-of-a-coin) for no changes, this result is unlikely. The two-tail P value here was 0.0026 which is statistically significant.

The Net Conversion Rate Sign Test result was consistent with the effects mentioned above. The experiment net conversion rate was higher than the control group on 10 of the 23 days. The two-tail P value was 0.6776 which is not statistically significant.

Summary

I did not use the Bonferroni correction in my experiment since the Gross Conversion and Net Conversion rates are dependent. An increase in enrollments will likely raise the net conversion (enrollments + payments) rate and vice versa. The size hypothesis test and sign test both came to the same conclusions:

1) The experiment resulted in a statistically significant change in Gross Conversion. The alternative hypothesis was correct.
2) The change in Net Conversion was **not** statistically significant. The null hypothesis is correct.

Recommendation

This experiment shows that the User Interface change does meet its goals. Enrollments were reduced by 2% but the net conversion rate fell by a much smaller number. Clearly not all the lost enrollments are affecting the paying customers. We can guess that many of those lost enrollments were users who may have later become frustrated and cancelled their enrollments before that first payment.

Despite not being a significant change, the experiment group did show an almost 0.5% lower net conversion rate than the control group. The change may be due to chance or it may be repeatable in further experiments.

If Udacity's business would be largely affected by even a 0.5% reduction in net conversions, then perhaps more experiments are needed here. Gathering new sample data would only take us 18 more days..

If Udacity's support and billing team is being overwhelmed with frustrated users requesting refunds or cancelling their enrollments, perhaps the potential small impact on net conversion would not be so big of a deal.

Implementing the change for all users would best be a decision made by Udacity's business teams after evaluating the results of the experiment. I call this a judgement call.

Follow-Up Experiment

I propose a follow-up experiment that eliminates 'frustrated users' entirely.

Instead of prompting users for payment details when they start their free trial, I propose moving this to the end of the trial. "Start a free trial (no credit card needed!)" sounds enticing and has no risk. After the end of the trial, a user will be shown a screen displaying their course progress and inviting them to continue with their coursework by submitting their payment details. Only users who are interested in continuing the course will be charged.

This change will likely reduce the rate of free trial enrollments that become paying customers. On the other hand, from the above experiment, we noticed only ~8% of cookies that visit the Udacity page try out the free trial. With a no-risk (no credit card) free trial, the enrollment rate may increase and make up for the reduced ratio of paying customers.

This experiment would measure the same metrics as the above experiment. The primary difference would be that Click-through-probability would become an Evaluation Metric as it will likely change due to the changes to the "Start Free Trial" button. The unit of diversion will stay the same. We would be tracking cookies as they transition from anonymous visits to enrolled (and paying) users.

Proposed Change: Modify the 'free trial' so that no credit card is needed. When the free trial is over, the user should be presented with an invitation to begin paying for the course. Students' progress from the free trial will be saved. Note: Students should have access to the reviews of already submitted projects but will not be able to resubmit until they become paying users.

Hypothesis: Students who enjoy the class and coursework will pay and continue to use the service. Users who lose interest in the course and fail to return to the page will never be charged (and will not become 'frustrated'). I also expect that a free trial with no strings attached (no credit card needed) will translate into more enrollments.

Unit of Diversion: The unit of diversion is a cookie. This experiment must track cookies and their status as unregistered, enrolled in a free trial, or paying customer.

Metrics:

There will be one invariant metric:

> **Number of Cookies**: Both control and experiment groups should be assigned the same number of cookies.

There are four evaluation metrics:

> **Click-Through-Probability**: The rate of "Start Free Trial" clicks may change (increase?) if a credit card is no longer required.
>
> **Gross Conversion**: Changes to click-through-probability will also affect gross conversion (the number of users enrolling in a free trial) since no credit card is required.
>
> **Retention**: Users (cookies) now decide whether to transition from a free trial to a paying user. This rate may also change in the experiment group.
>
> **Net conversion**: Due to Gross Conversion changes, it is likely that Net Conversion will also change.

Similar to the above test, I would propose that 50% of our traffic be diverted into an experiment group and 50% in a control group. The changes in this experiment would likely be much greater than the last experiment largely because I expect that the users clicking “Start Free Trial” will increase along with the number of students who enroll in the course.