Overview

Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available 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.This screenshot shows what the experiment looks like.

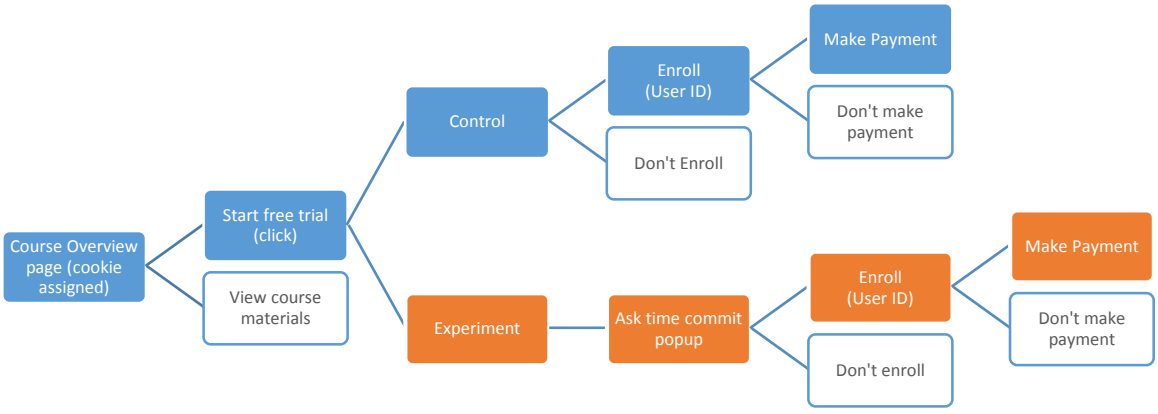
The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete 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 unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

Experiment Design

Overview

There are two options for users on the course overview page – start a free trial or view course materials. This experiment is designed to test changes on the start free trial option in order to improve the student experience. The following diagram indicates the path of the user and where the experimental condition is introduced. Items that are measured are in a solid color, and the experimental condition is indicated in orange.



Metric Choice

Invariant Metrics

Cookies and clicks are measured before the experimental condition is introduced (in orange). One expects that these metrics will have a statistically equivalent distribution between the experiment and control group in a properly designed and executed experiment. Click through probability is also an important metric as it measures the total impact, and it is calculated on cookies and clicks. These three metrics will be used to do a sanity check before the data is analyzed.

Metric	Description	d_{min}	Formula
Cookies	Number of unique cookies to view the course overview page	3000	Count
Clicks	Number of unique cookies to click the "Start free trial" button	240	Count
Click through Probability	That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.	0.01	= Clicks / Cookies

Evaluation Metrics

Evaluation metrics were first screened based on their relevance to the hypothesis. Number of user-ids and Gross conversion both measure the first condition, which is diverting students who can't make the time commitment to have a good chance of success.

Retention and Net conversion both measure the second condition of the hypothesis, which is not changing the 14 day enrollment probability.

However, it is desirable to use the minimum number of evaluation metrics to avoid false positives (i.e. measure a difference that is not real). Solving for this either requires a lower α for each individual test to maintain $\alpha_{overall}$ or using the Bonferroni correction (which also lowers the α)

Gross conversion and Net conversion are ultimately analyzed for the experiment. These two metrics normalize for changes in traffic by using unique cookies as their denominator. This removes other variables that may affect the number of user-ids, such as changing traffic patterns.

Metric	Description	d_{min}	Formula
Gross conversion	Number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button	0.01	= enroll / clicks (Oct 11 to Nov 2)
Net conversion	number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button	0.075	= payment / clicks (Oct 11 to Nov 2)

There are two requirements defined in the experiment hypothesis.

Gross conversions will change

$$H_0: p_{exp} - p_{cont} = 0, H_a: p_{exp} - p_{cont} \neq 0$$

Net conversions will not change

H_0: $p_{exp} - p_{cont} \neq 0$, H_a: $p_{exp} - p_{cont} = 0$

In addition, practical significance boundaries are established for each metric to set a minimum change to make launching the experiment worthwhile. The matrix below summarizes the decision criteria for launching the experiment where both the gross conversions and the net conversions must pass.

Decision matrix			
Gross conversions	Pass	Fail	Fail
Net conversion	Fail	Pass	Fail
$d = p_{exp} - p_{control}$		$-d_{min}$ 0 d_{min}	

Unused Metrics

Ultimately, the number of user ids was discarded because there were other variables not related to the experiment that could affect this variable (i.e. a change in traffic)

Retention was discarded because it drives an extra-ordinarily high sample size to get statistically significant results which would have added months to the experiment.

- **Number of user-ids:** That is, number of users who enroll in the free trial. (d_{min}=50)
- **Retention:** That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. (d_{min}=0.01)

Measuring Standard Deviation

metric	Baseline (N)	Formula	Sample size (N)	se (sample size N)
Unique cookies to view page per day = Cookies:	N = 40000	Count	5000	
Unique cookies to click "Start free trial" per day = Click:	N = 3200	Count	400	
Enrollments per day:	N = 660	Count		
Probability of payment, given enroll	p = 0.53	Enroll * p		
Probability of enrolling, given click:(Gross Conversion)	p = 0.20625	Enroll / click	400	0.0202
Probability of payment, given click (Net Conversion)	p = 0.1093125	Payment / click	400	0.0156

Fortunately, both of the metrics selected for evaluation are probabilities, and these tend to be binomial for large sample sizes. For this reason, the analytic estimate is expected to be similar to the empirical estimate of variability.

Sizing

Number of Samples vs. Power

Sample sizes (pageviews) were calculated for both metrics, using α = 0.05 and α = 0.025 (Bonferroni correction).

	Input	Minimum detectable effect d_{min}	Samples / group	Total samples required
α = 0.05, ß= 80%, http://www.evanmiller.org/ab-testing/sample-size.html				
Gross conversions	p = 20.6%	1%	= 25,812 / 0.08	645,300
Net conversion	p = 10.9%	0.75%	= 27,345 / 0.08	683,625
α = 0.05, ß= 80%, N = 5000, 2016 TDX_calc_N.R				
Gross conversions	SE = 0.0202	1%	25,710	642,750
Net conversion	SE = 0.0156	0.75%	27,210	680,250
α = 0.025, ß= 80%, N= 5000, 2016 TDX_calc_N.R				
Gross conversions	SE = 0.0202	1%	31,100	777,500
Net conversion	SE = 0.0156	0.75%	32,910	822,750

Net conversions with α = 0.05 is used for the experiment for a handful of reasons. First, it is the metric with the highest # of samples require, so it will allow sufficient samples for all metrics. Second, # of pageviews is high relative to the existing traffic to the site and using the Bonferroni correction will result in a 20% longer test. In addition, the conversions are not independent of each other and using the more stringent Bonferroni correction is overly conservative for this test.

Duration vs. Exposure

This experiment requires some hard trade-offs between duration and exposure. The number of pageviews required compared to the daily traffic means that 17 days is the minimum duration assuming all traffic is diverted for the test. In addition, there is a 2 week delay to finalize results because net conversions cannot be measured until 14 after the initial click.

Opportunity Cost has to be considered for increasing the duration, and includes factors such as

Prevents running other tests for different features that may also improve the site.

It also delays the launch of a potentially valuable feature.

Counter to the duration is the exposure - diverting a large portion of traffic for a feature that has not been fully vetted runs the risk of a poor user experience. (This can be addressed with more robust testing to ensure device, OS, and other compatibility is acceptable).

Ultimately, the recommendation is to divert 100% of the traffic which results in an 18 day experiment.

Experiment Analysis

Sanity Checks

All Sanity checks pass at 95% CI

	Baseline	Lower	Upper	Observed	Result
Cookies	p = 0.5	0.4988	0.5012	0.5006	Pass
Clicks	p = 0.5	0.4959	0.5041	0.5005	Pass
Click through probability	p_{pool} = 0.0822 d = $p_{exp} - p_{cont}$ = 0	-0.0013	0.0013	0.0001	Pass

Result Analysis

Effect Size Tests

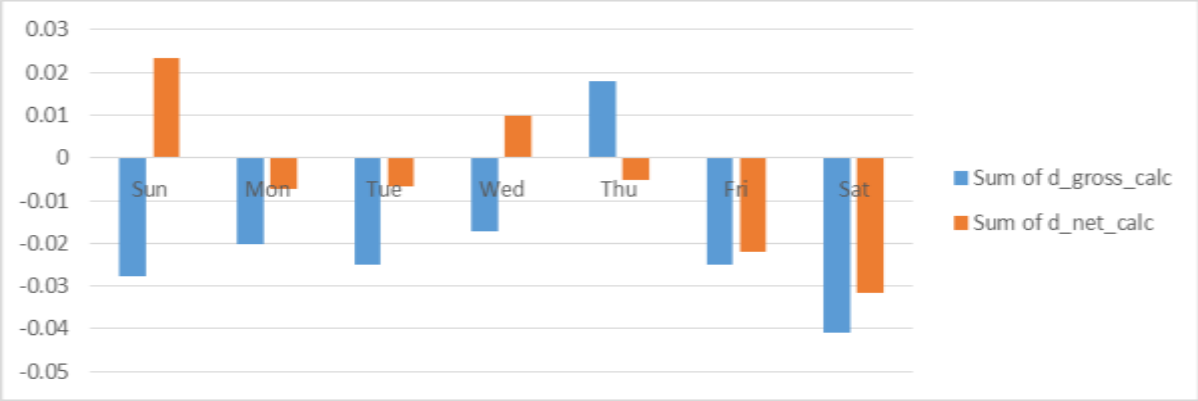
All Effect Size Tests pass the requirements at 95% Confidence Interval:

Metric	d = p_{exp} -p_{cont}	d_{min}	Lower	Upper	Result
Gross Conversions	-0.0206	0.01	-0.0291	-0.0129	Pass
Net Conversions	-0.0049	0.0075	-0.0116	0.0019	Pass

Decision matrix with results			
Gross conversions	Pass -0.0291 -0.026 -0.0129	Fail	Fail
Net conversion	Fail	Pass -0.0116 -0.0049 0.0119	Fail
d = p_{exp} - $p_{control}$		-d_{min} 0	d_{min}

For net conversions, there is some overlap between the confidence interval and the fail condition. This should be highlighted to the decision maker as it is definitely not desirable to decrease the level at which students convert!

One strategy to address this is to dig deeper into the test results. There is a noticeable change by the day of the week between $d_{Gross\ conversions}$ and $d_{Net\ conversion}$. Notably, the difference between the experimental and control group is negative and greater than d_{min} for Friday and Saturday. This suggests an issue with the experiment, or perhaps a change in user behavior based on proximity to the weekend!



Sign Tests

Metric	Count d<0	Count total	p	Result
Gross conversions	19	23	0.0026	Significant (pass)
Net conversion	13	23	0.6776	Not significant (pass)

Summary

Two methods were utilized to analyze the experimental results.

First the effect size test was utilized to confirm a statistical difference between the gross conversion probability and the net conversion probability between the control and experimental group. Bonferroni's correction was not used. The decision was made to keep the duration at 35 days and increase $\alpha_{overall}$ from 0.05 to 0.0975.

The sign test confirmed the effect size test – it is important to note that the net conversion result of not significant is actually a confirmation of the test parameters to not change the net conversion probability.

Recommendation

My conclusion is that this experiment successfully shows the improvement indicated by the hypothesis to a statistical and practical significance. Given the overlap between the confidence interval for the net conversions and the practical boundary, I would propose an additional experiment to the decision makers to better understand the change in net conversions by day.

Follow-Up Experiment

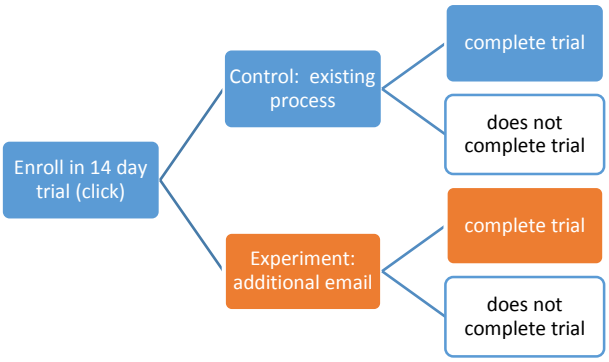
Background

The previous experiment indicates that Udacity can properly filter students that have a better probability of success by guiding them to an enrolled status if they expect to be able to devote 5 or more hours / week and a "audit" status (courses, but no coaching) otherwise. This allows a better user experience for the students who presumably have a better chance of success.

The experiment addresses an anomaly from this same experiment that indicates there is a noticeably better net conversion for students that enroll between Sun –Thurs and lower net conversion for students that enroll on Fri-Sat.

Experiment overview

The experiment is to add an additional contact step to students focusing on the coaching feature, and the exact details will be based on a review of the logs of user-ids in the previous study (i.e. if most students that enroll but don't log in again for the 2 week trial after the 3rd day, an email will be sent at 2 days encouraging them to log in and contact a coach to continue with their course work).



The null hypothesis (H_0) is that there is no difference between the probability of the control and the experimental group to make a payment ($p_{exp} – p_{cont} = 0$)

The alternate hypothesis (H_A) is there is a statistically significant difference between the probability of the control and experimental group to make a payment ($p_{exp} – p_{cont} \neq 0$)

The unit of diversion is a click to start the trial, although user id will be used to track activity.

The metrics will be similar to the previous experiment.

Invariant metrics:

- clicks to enroll in a 14 day trial
- count of sign ins before intervention
- probability of user-ids logging in at least on additional time before intervention (count of unique user ids / click.

Evaluation metrics

- retention = payments / clicks ($d_{min} = 0.075$)

Intuition that was built in the previous experiment is informing many of these decisions. For example, enrollments (User IDs) has too small of a population to provide results in a reasonable time period and so this is not selected as an evaluation metric.

Using similar metrics allows for comparison with the previous experiment, and having more than one invariant metrics allows additional sanity checks.

Resources

https://discussions.udacity.com/t/final-project-calculating-standard-deviations/27356/19
http://www.evanmiller.org/ab-testing/sample-size.html
https://discussions.udacity.com/t/sanity-checks-confidence-intervals-observed-values/27175/22