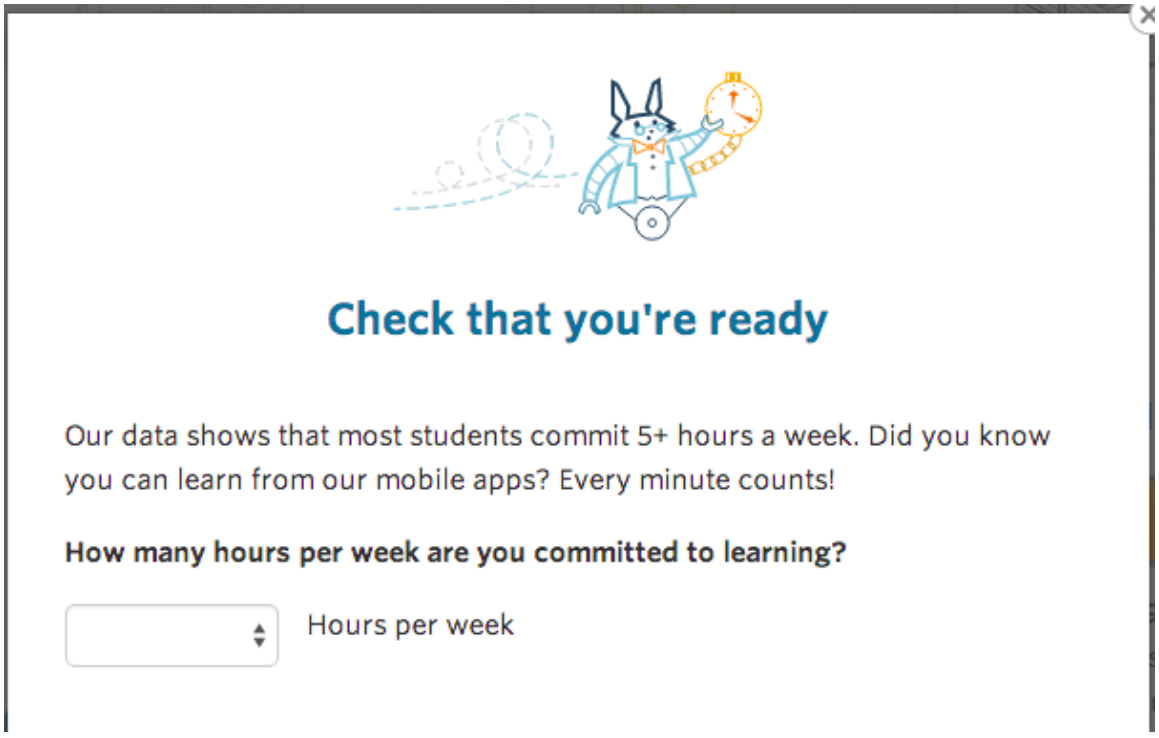
Author: Navina Govindaraj
Date: May 2017

Experiment Overview: Free Trial Screener

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. The below 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 leave 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

Metric Choice

Invariant Metrics	Number of cookies, Number of clicks
Evaluation Metrics	Gross conversion, Net conversion

• **Number of cookies:** That is, number of unique cookies to view the course overview page. This is a good invariant metric since cookies are loaded before the experiment and are hence are independent of the experiment.

• **Number of clicks:** That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). This is a good invariant metric because the clicks happen before the experiment is shown.

• **Gross conversion:** That is, 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. This is a good evaluation metric since it is dependent on the experiment (which is why it cannot be a good invariant metric) and also it shows the probability to succeed. This metric is expected to be lower for the experiment group than the control group because there will be users who won't be certain if they can commit more than 5 hours per week. Users in the control group will go ahead and enroll since the pop-up message is not being shown to them.

• **Net conversion:** That is, 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. This is a good evaluation metric since this is dependent on the experiment (hence not a good invariant metric). The experiment is being conducted based on the premise that users who are aware of the time that needs to be put into the program before clicking on the free trial, have better chances of sticking with the course beyond the trial period. It would be interesting to see if the results of the experiment line up with the expectation.

• **Number of user-IDs:** That is, number of users who enroll in the free trial. This is neither a good invariant metric nor a good evaluation metric. The number of user-IDs (number of enrollments) could fluctuate a lot on a given day. Gross conversion, which is the number of user-IDs divided by the number of 'start free trial' clicks is a better metric.

• **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. This is a good invariant metric because the clicks happen before the user sees the experiment, and are thus independent of the experiment. But this has not been chosen as an invariant metric for the experiment because both number of cookies and number of clicks are anyway being used as invariants and choosing this would be redundant.

• **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. This is a good evaluation metric because it is dependent on the experiment (hence not a good invariant metric). Retention is expected to be higher in the experiment group since this group has users who are aware of the kind of time commitment required, unlike users in the control group whose chances of dropping out are higher since they are unaware of the time commitment. However, this metric has been disregarded because it would take too long to measure gather the required data. Further elaboration has been provided under 'Sizing'.

Launch criteria

Gross Conversion	should decrease significantly (cost impact)
Net Conversion	should not decrease significantly (revenue impact)

Table of Baseline Values

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

Measuring Variability

List the standard deviation of each of your evaluation metrics.

Metric	p	N	Standard Deviation
Gross conversion	0.20625 (Probability of enrolling, given click)	5000 * 0.08 = 400	$\sqrt{\frac{0.20625*(1-0.20625)}{400}} = 0.0202$
Net conversion	0.1093125 (Probability of payment, given click)	5000 * 0.08 = 400	$\sqrt{\frac{0.1093125*(1-0.1093125)}{400}} = 0.0156$

- Since the unit of analysis is ‘cookie’ in both gross conversion and net conversion, and the unit of diversion is the same, the analytical estimate is likely to be accurate.

Sizing

Number of page views needed to power the experiment:

- The Bonferroni correction will not be used in this analysis.
  Using this calculator

alpha = 0.05, beta = 0.2

Evaluation Metric	Baseline Conversion Rate	Minimum observable effect (d_{min})	Required Sample Size	Required Page views (control + experiment groups)

Gross conversion	20.625%	1%	25,835	$\frac{25,835*40,000}{3200} = 322,937.5 * 2 = 645,875$
Net conversion	10.931%	0.75%	27,413	$\frac{27,413*40,000}{3200} = 342,662.5 * 2 = 685,325$
Retention	53%	1%	39,115	4,741,212

The retention metric is being dropped as an evaluation metric because it would take 117 days to gather data, which is too long for an A/B test.

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. What percentage of Udacity's traffic would you divert to this experiment (assuming there were no other experiments you wanted to run simultaneously)? How risky do you think this experiment would be for Udacity?

- With the traffic split 65-35 between the experiment and control groups, this would mean that the percentage of Udacity's traffic that would be diverted to this experiment would be 0.65
- A daily traffic of 40,000 unique cookies, would result in 40,000*0.65 = 26,000 diverted to the experiment
- Days required to run the experiment = $\frac{685,235}{26,000} = 26.36$ = 27 days
- The change being tested is not very risky, there is no issue with regard to sensitivity of the data being used for experimentation, and no harm can come to anyone as a result of conducting the experiment. The duration of 27 days is a reasonable period to run the test.

EXPERIMENT ANALYSIS

Sanity Checks

For each of the invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check.

Invariant Metric	Number of cookies	Number of clicks
Control + Experiment Group Total	345,543 + 344,660 = 690,203	28,378 + 28,325 = 56,703
Std.Error	$\sqrt{\frac{0.65*0.35}{690,203}} = 0.00057$	$\sqrt{\frac{0.65*0.35}{56,703}} = 0.002$
Margin of Error	1.96*0.00057 = 0.0012	1.96*0.002 = 0.004
Lower Bound	0.5 - 0.0012 = 0.4988	0.5 - 0.004 = 0.496
Upper Bound	0.5 + 0.0012 = 0.5012	0.5 + 0.004 = 0.504
Observed Value	$\frac{345,543}{690,203} = 0.5006$	Yes
Passes	$\frac{28,378}{56,703} = 0.5005$	Yes

Result Analysis

The calculations in this section are based on the experiment results

Effect Size Tests

For each of the evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant.

Evaluation Metric	Gross conversion	Net conversion
p	$\frac{3785+3423}{17293+17260} = 0.2086$	$\frac{2033+1945}{17293+17260} = 0.1151$
Std.Error (SE)	$\sqrt{\frac{0.2086*0.7914}{(1/17293)+(1/17260)}} = 0.00437$	$\sqrt{\frac{0.1151*0.8849}{(1/17293+1/17260)}} = 0.0034338$
Margin of Error	1.96*SE = 0.0085652	1.96*SE = .0067518
Mean	$\frac{3423}{17260} - \frac{3785}{17293} = -0.0205549$	$\frac{1945}{17260} - \frac{2033}{17293} = -0.0048737$
Lower Bound	-0.0291	-0.0116
Upper Bound	-0.0120	0.0019
Statistical Significance	Yes	Yes
Practical Significance	No	No

Sign Tests

For each evaluation metric, do a sign test using the day-by-day breakdown. If the sign test does not agree with the confidence interval for the difference, see if you can figure out why.

Evaluation Metric	Number of "successes" observed	Number of trials or experiments	p-value	Statistical Significance
Gross conversion	4	23	0.0026	Yes
Net conversion	13	23	0.6776	No

Summary

Bonferroni correction has not been used in the analysis. In this case, multiple (two) metrics have to meet the criteria at the same time for us to launch the change. Bonferroni correction reduces the statistical significance level and is hence very conservative. This increases the chance of rejection of one of the multiple metrics, which in turn means the rejection of the entire proposal. As a result, the probability of Type II error goes up. Bonferroni correction is more appropriate in situations where we have multiple metrics and when the confirmation of ANY of them is sufficient for launch. Here, it is justified to be conservative in order to reduce the probability of Type I error.

Recommendation

The recommendation is not to launch the change of including the time commitment message. Gross conversion will decline, meaning that there would be a significant reduction in enrollments that do

not pay (thereby reducing cost). However, net conversion is neither statistically nor practically significant, which means that the overall paid enrollments will not see an increase (negative impact on revenue). Though this outcome matches the second launch criteria, the CI of net conversion indicates a propensity to decrease. This is not in line with the business objective to increase paid users, and hence this feature cannot be implemented.

Follow-Up Experiment

If you wanted to reduce the number of frustrated students who cancel early in the course, what experiment would you try? Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

Announce an incentive package to users who already have a free account on Udacity, but haven't taken the leap to paying for any Nanodegree program, that if they accept the challenge of completing one Nanodegree course in one month (starting immediately), they will receive a 50% waiver on the cost for the entire program (redeemed on completion of the Nanodegree).

Null hypothesis: There will be no change in net conversion between control and experiment groups

Unit of diversion: user-ID, since only users who have created a free account will be able to see the experiment and this will ensure that a signed in user is not part of both experiment and control groups

Invariant metric: Number of user-IDs. Since only half the logged-in users(experiment group) will be seeing the incentive announcement, the number of users logging into the website will not change and is therefore independent of the experiment.

Evaluation metric: Retention, that is, number of user-IDs to complete a course within the 30-day boundary (and thus make at least one payment) divided by the total number of unique user-IDs to click the incentive program message button.

This would ensure that all motivated students already under the free membership pay for the program sooner, thereby helping them commit to the program. This would also keep students who are not yet ready to put in the time and effort to not pay and then get frustrated realizing the sort of commitment required.
