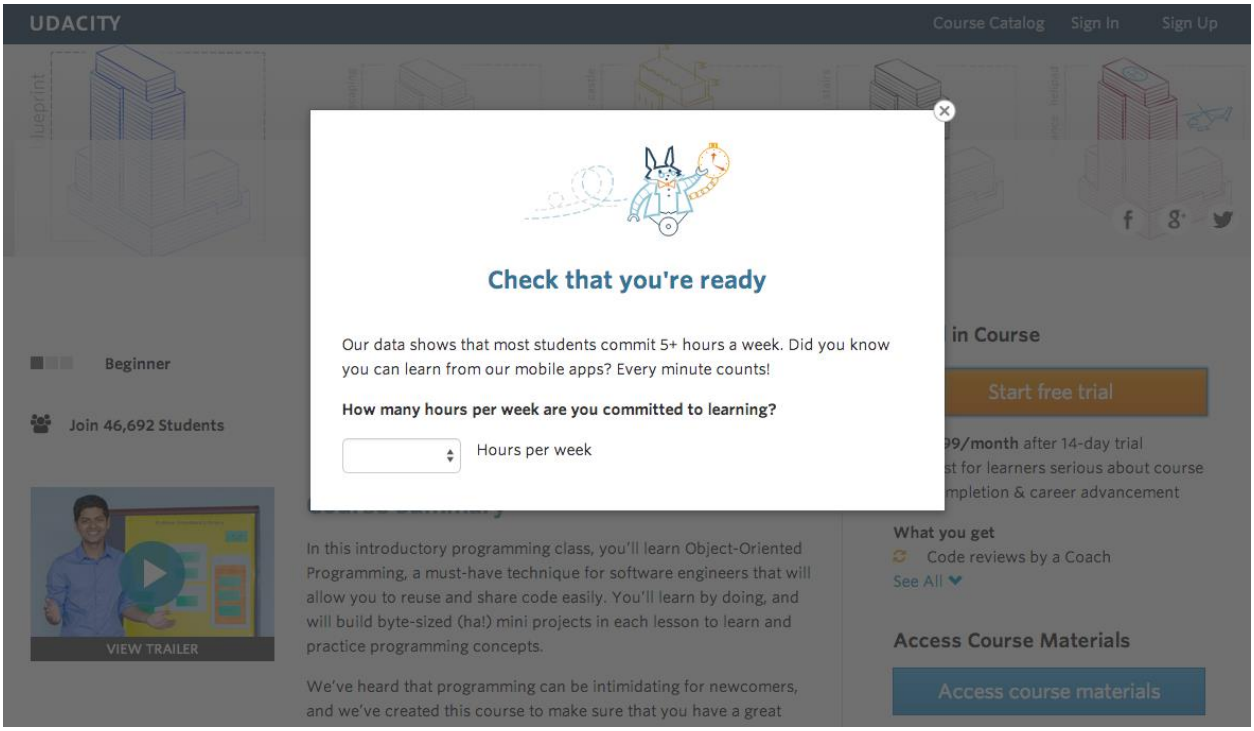
Experiment Overview: Free Trial Screener

Background:
At the time of this experiment, Udacity courses currently have two options on the course overview 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 is that by setting clearer expectations for students upfront, this screener will reduce 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 students enroll 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

The practical significance boundary for each metric, that is, the difference that would have to be observed before that was a meaningful change for the business, is given in parentheses. All practical significance boundaries are given as absolute changes.

Any place "unique cookies" are mentioned, the uniqueness is determined by day. (That is, the same cookie visiting on different days would be counted twice.) User-ids are automatically unique since the site does not allow the same user-id to enroll twice.

Invariant Metrics:

- **Number of Cookies**: That is, number of unique cookies to view the course overview page. (dmin = 3000)
- **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). (dmin = 240)
- **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. (dmin = 0.01)

Invariant metrics are used for sanity checks, so the value of these metrics shouldn’t be affected by the implementation of this new feature. The new layer that Udacity added to the user funnel, is after users click on the start free trial button. Therefore, the metrics that are related to the customer journey before reaching to this button should remain invariant.

Evaluation Metrics:

- **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. (dmin = 0.01)
- **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. (dmin = 0.01)
- **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. (dmin=0.0075)

In order to launch this new feature, I am expecting to see a rise in retention meaning that more students stay after the free trial in the experiment group, a decrease in gross conversion meaning that there are less frustrated students in the experiment group, and not a big change in net conversion meaning that we don’t lose a significant number of users who make payments in the experiment group.

Measuring Variability (Standard Deviation)

For each of the evaluation metric, I will provide an analytic estimate of its standard deviation, given a sample size of 5000 cookies visiting the course overview page. Data is enclosed in the table of baseline values.

	P(Probability)	N (sample size)	SE (standard error)
Gross conversion	660/3200=0.20625	5000*0.08=400	0.020230604
Retention	0.53	5000*(660/40000) = 82.5	0.054949012
Net conversion	660*0.53/3200 = 0.1093125	5000*0.08=400	0.015601545

Sizing

Number of Samples vs. Power

Given alpha = 0.05 and beta = 0.2, the sample size per variation is calculated through the link.
I will not use the Bonferroni correction during my analysis phase.

	Baseline Rate	Minimum Detectable Effect	Sample size per variation	How many pageviews will be needed?
Gross Conversion	20.625%	0.01	25,835	25,835*2 / (3200/40000) = 645,875
Retention	53%	0.01	39,115	39,115*2 / (660/40000) = 4741212.121
Net Conversion	10.93125%	0.0075	27,413	27,413*2 / 0.08 = 685325

In order to power my experiment appropriately, I will need 4741212 pageviews.

Duration vs. Exposure

	Sample size per variation	How many pageviews will be needed?	Required Days (if 100% traffic is exposed)
Gross conversion	25,835	645875	16.146875
Retention	39,115	4741212.121	118.530303
Net conversion	27,413	685325	17.133125

For this test, I need to choose the largest number in the last column of the table above, that is it will take 119 days to get required number of cookies for retention metric. This is a very long time and it's too costly. Then we move on to the second largest number, that is 17 days to acquire enough number of cookies for net conversion metric. Please note that the calculation of required days assumes that 100% traffic is exposed to this experiment, which is not realistic and also too risky. Assume that 80% of Udacity's traffic is diverted to this experiment, we will need to run this experiment around 21 days.

Experiment Analysis

Sanity Checks

For each of my invariant metrics, I will give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. The experiment data can be found in this spreadsheet.

In order to perform sanity check for number of cookies and number of clicks on "Start free trial" (count). I will calculate a confidence interval around the fraction of events I expect to be assigned to the control group. And the observed value will be the actual fraction that was assigned to the control group. For invariant metrics, I expect equal diversion into the experiment and control group, so I am going to check if the difference is random or significant. I can model the assignment of cookies to control and test group by a binomial distribution, with 50% probability of success (success is defined as being assigned to the control group).

	Pageviews	Clicks
Standard Error	(0.5*0.5/(345543+344660))^0.5= 0.000601841	(0.5*0.5/(28378+28325))^0.5= 0.002099747
Margin of error = SE*Z-score(alpha=0.05)	1.96*0.000601841 = 0.001179608	1.96*0.002099747 = 0.004115504
Lower Bound = p-margin of error	0.5-0.001179608 = 0.498820392	0.5-0.004115504 = 0.495884496
Upper Bound = p+margin of error	0.5+0.001179608= 0.501179608	0.5+0.004115504 = 0.504115504

	Lower Bound	Upper Bound	Observed	Does the Metric pass the sanity check?
Number of cookies	0.4988	0.5012	0.50064	YES
Number of clicks on "Start free trial"	0.4959	0.5041	0.5005	YES

Since the observed ratios of cookies and clicks in control group falls within the confidence interval, the differences are not significant. Number of cookies and number of clicks on "Start free trial" passes the sanity check.

In order to perform sanity check for click-through-probability on “Start free trial” (probability), I will construct a confidence interval for a difference in proportions, then check whether the difference between group values falls within that confidence interval. In this case, I expect the difference between click-through-probability of the two groups to be zero.

	CTR
Control	28378/345543 = 0.082125814
Experiment	28325/344660 = 0.082182441
Observed Difference	0.082182441 - 0.082125814 = 5.66271E-05
p	0
Pooled Probability of a click	56703/690203=0.082154091
SEpool	(0.082154091*(1-0.082154091)*(1/345543+1/344660))^0.5 = 0.000661061
Margin of error = SE * 1.96 (a=0.05)	0.000661061*1.96 = 0.001295679
Lower Bound = p-margin of error	0-0.001295679 = -0.001295679
Upper Bound = p+margin of error	0+0.001295679 = 0.001295679

	Lower Bound	Upper Bound	Observed	Does the Metric pass the sanity check?
Click-through-probability on “Start free trial”	-0.0013	0.0013	0.0001	YES

Since the observed Click-through-probability on “Start free trial” in control group falls within the confidence interval, the difference is not significant. This metric also passes the sanity check.

Now that all the sanity check passed, I will proceed to the rest of the analysis.

Result Analysis

Effect Size Tests

For each of my evaluation metrics, I will construct a 95% confidence interval around the difference between the experiment and control group, and check whether each metric is statistically and practically significant.
A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident that there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident that there is a change that matters to the business.)

	Gross Conversion	Retention	Net Conversion
Pooled Probability	0.208607067	0.551886792	0.115127485
Pooled Standard Error	0.004371675	0.01172978	0.003434134

Marginal Error = Pooled Standard Error*1.96	0.008568484	0.022990369	0.006730902
d_hat= observed difference between control and experiment	-0.020554875	0.031094805	-0.004873723
Lower bound = d_hat - marginal error	-0.029123358	0.008104436	-0.011604624
Upper bound = d_hat + marginal error	-0.011986391	0.054085174	0.001857179
dmin	0.01	0.01	0.0075
Statistical Significance	YES	YES	NO
Practical Significance	YES	NO	NO

Notes:
The enrollment and payment data are not complete. When doing the calculations, I will filter out the days with incomplete information.
I didn't use the Bonferroni correction here because the three evaluation metrics are closely related to each other and using Bonferroni correction would be too conservative in this case. Applying the Bonferroni correction means that the alpha level for each metric would be around 1.67% instead of 5%, and the confidence intervals would be much wider.

For Gross Conversion, the difference between control and experiment group is both statistically significant and practically significant as the confidence interval doesn't include 0 or practical significance.
For Retention, the confidence level doesn't include 0, but include the practical significance, therefore the difference between control and experiment is statistically significant, but not practically significant.
For Net Conversion, the difference between control and experiment group is insignificant since the confidence interval include both zero and negative dmin.

Sign Tests

The goal of sign tests is to check whether the signs of the difference of the metrics between the experiment and control groups agree with the confidence interval of the difference. For each of my evaluation metrics, I will conduct a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. The calculation is done through the online p-value calculator.

Gross Conversion:
Count of days where Gross Conversion is higher in the experiment group than that in the control group is 4(out of 23 days), therefore the p-value is 0.0026. Since 0.0026 is smaller than 0.05, gross conversion is significant at an alpha=0.05 level.
Retention:
Count of days where Retention is higher in the experiment group than that in the control group is 13(out of 23 days), therefore the p-value is 0.6776. Since 0.6776 is larger than 0.05, retention is not significant at an alpha=0.05 level.
Net Conversion:
Count of days where Net Conversion is higher in the experiment group than that in the control group is 10(out of 23 days), therefore the p-value is 0.6776. Since 0.6776 is larger than 0.05, net conversion is not significant at an alpha=0.05 level.

Summary

The sign test agrees with the hypothesis test on the effect size of gross conversion and net conversion. The only discrepancy between the effect hypothesis tests and the sign tests is that for retention metric, the hypothesis test on the effect size showed statistically significant result, but the sign test didn't. Sign test has lower power than the effect size test, which is frequently the case for nonparametric tests, and that's the price you pay when you are not making any assumptions. This is not necessarily a red flag, but it's worth digging deeper and see what's going on. Looking at the day-by-day breakdown again, all Sundays were positive. The change may have smaller or no effect during the week, but a relatively larger effect on Sundays. We might need to dig deeper into why the change didn't affect weekday visitors to have an idea on how to iterate on the change to help it affect more users. However, since this gap between weekday and weekend users are not too obvious, what we can also do is to break up the traffic into different demographic groups and see if there is a specific nationality, age or education group is driving this difference. We can also break down the traffic into different platform users such as mobile apps or websites.

I didn't use the Bonferroni correction for any calculations because it would be too conservative in the case of highly related metrics.

Recommendation

I will not launch this feature. First of all, only gross conversion is both statistically and practically significant with a negative value, which means we are 95% positive that gross conversion will decrease significantly. However, we are unable to conclude if this feature will increase the retention rate overall. The net conversion rate is not statistically significant nor practically significant, meaning that we are 95% sure that this feature will not change this value. Hypothesis test of retention shows that retention is statistically significant, but is not significant enough to make a business impact. In order to understand the retention, we need to run additional experiments and tests to determine the change in retention.

Gross conversion is the only metric that is both statistically significant and practically significant. This value decreased in the experiment group, just as we expected. However, the results of net conversion were not significant. This means that we are not confident if the free trial screener will change this metric and we are not sure how the number of users who make payments change. Although the decrease in gross conversion will probably make the number of frustrated students decrease too, but it's not clear how the number of students who pass the free trial period and make payments change.

Follow-Up Experiment: How to Reduce Early Cancellations

In order to reduce the number of frustrated students who cancel early in the course, I have designed the following experiment. I want to add a feature of pop-up motivation message once students enroll in the free trial.

Hypothesis:

My hypothesis is that adding a pop-up motivation message would push some students who might otherwise drop out during the 14-day trial to continue and even finish the course.

Unit of Diversion:
User-id.
Metrics:
The invariant metrics in this experiment will be the same as the previous experiment, that are number of cookies, number of clicks and click-through-probability. Since this feature will be added after the click on the “Start free trial” button, these metrics shouldn’t be affected and should stay unchanged.
The evaluation metric will be retention rate, 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.