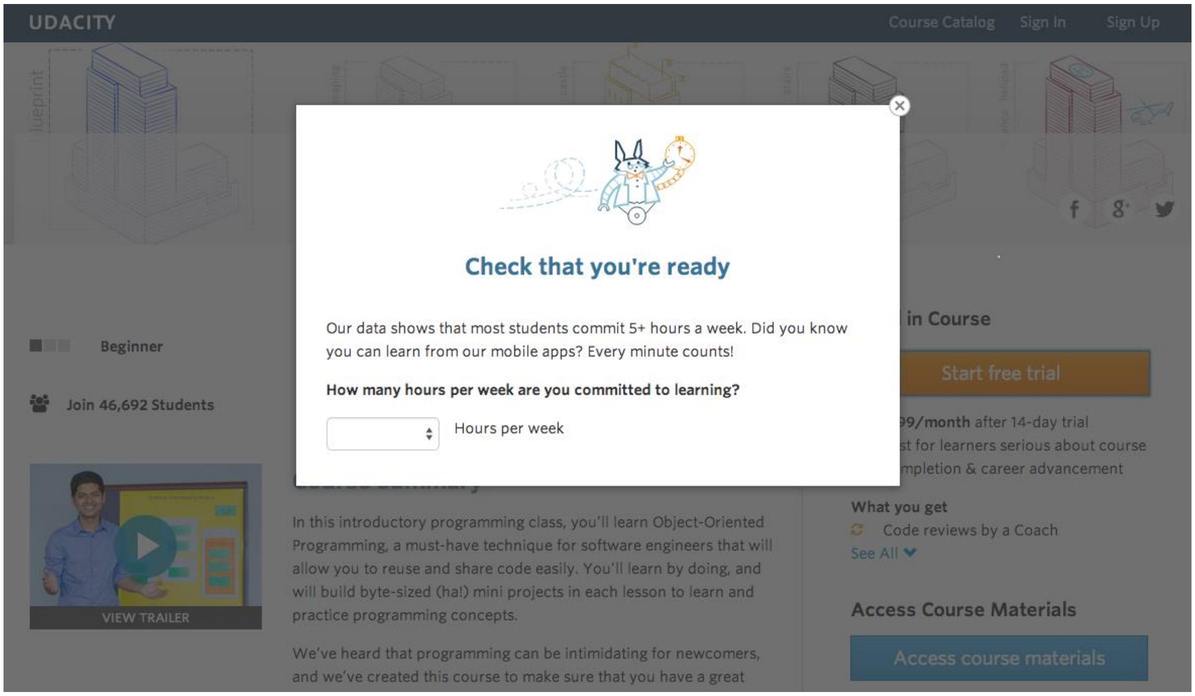
P7: Design an A/B Test

Experiment Design: Free Trial Screener

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.



Check that you're ready

Our data shows that most students commit 5+ hours a week. Did you know you can learn from our mobile apps? Every minute counts!

How many hours per week are you committed to learning?

Hours per week

In this introductory programming class, you'll learn Object-Oriented Programming, a must-have technique for software engineers that will allow you to reuse and share code easily. You'll learn by doing, and will build byte-sized (ha!) mini projects in each lesson to learn and practice programming concepts.

We've heard that programming can be intimidating for newcomers, and we've created this course to make sure that you have a great

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.

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric.

Invariant Metrics: Number of cookies, Number of clicks, Click through probability
Evaluation Metrics: Gross Conversion, Net Conversion

Invariant Metrics are those metrics that remain same in the control and the experiment group.

Evaluation Metrics are those metrics that help make business decisions. The goal of conducting the experiment is to find out that if the experiment has affected the evaluation metrics i.e. the goal is to find that if the experiment has resulted in any statistically significant output.

The **hypothesis** of adding the free trial screener 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.

Evaluation Metrics are those metrics that may help measure what the experiment is trying to achieve. Here, evaluation metrics are those that help get a measure directly or indirectly of the students who left the free trial because they did not have enough time and of those students who continued past the free trial and eventually completed the course.

Number of cookies: That is, number of unique cookies to view the course overview page.

Number of cookies is to remain same in the control and the experiment group and hence, is an invariant metric.

Number of cookies is insufficient to measure the result of the experiment and cannot be used as an evaluation metric.

Number of user-ids: That is, number of users who enroll in the free trial. Enrollment happens after the free trial screener is triggered and hence, the enrollments will be different for the control group and the experiment group which is why it cannot be used as an invariant metric.

The number of user IDs can be used as an evaluation metric. However, it is not normalized. Hence, it could be an evaluation metric but not the best one. We refrain from using it in this case.

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). Number of clicks is to remain same in the control and the experiment group and hence, is an invariant metric as it happens before the free trial screener is triggered.

But it cannot be used to measure the effect of the experiment and hence, it cannot be used as an evaluation 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 happens before the screener and the metric remains same for both groups and hence, it is an invariant metric.

Since, it measures events that happen before the screener gets triggered, it cannot be used to measure the effect of the experiment and hence, it cannot be used as an evaluation metric.

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. The enrollment happens after the screener is triggered and hence, it cannot be a invariant metric.

Gross conversion helps measure the ratio of users enrolling to the users who clicked on start free trial i.e. the proportion of users who enrolled after seeing the screener. This is the evaluation metric as it captures the change in behavior between the control and experiment group as a result, of the screener.

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. The enrollment happens after the screener is triggered and hence, it cannot be a invariant metric.

However, it cannot be used as an evaluation metric because, it cannot capture the change in behavior due to the experiment.

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. The enrollment happens after the screener is triggered and hence, it cannot be a invariant metric.

However, it can be used as an evaluation metric because, it captures the change in behavior due to the experiment, at least it aims to.

Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

When reduction in Gross conversion is practically significant, and there is no significant change in net conversion due to the experiment (it is important that there is no reduction in this metric because that can result in decrease in revenue for the company) the experiment can be launched. However, it is important to consider the confidence intervals and the risk factor before launching.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

Standard deviation:

SE= sqrt(p(1-p)/N)

Gross conversion: SE = sqrt(0.20625*(1-0.20625)/(5000*0.08)) = 0.0202
Net conversion: SE = sqrt(0.1093125*(1-0.1093125)/(5000*0.08)) = 0.0156

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

Both the evaluation metrics, make use of number of cookies. Here, the unit of diversion is same as the unit of analysis. Therefore, the analytic estimate would be comparable to the the empirical estimate.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

No, I did not use Bonferroni correction during my analysis phase. It seems that the Bonferroni correction in our scenario is too conservative

Probability of enrolling, given click: 20.625% Baseline conversion rate
1% min d.
Samples needed: 25,835

Probability of payment, given click: 10.93125% Baseline conversion rate
0.75% min d.
Samples needed: 27,413 (chosen)

Ratio of page views to clicks = 0.08

Total pageview = 27413/0.08*2= 342662.5*2 = 685325

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. (These should be the answers from the "Choosing Duration and Exposure" quiz.) Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

The exposure can be determined by analyzing the risk involved. The screener is nothing but a reminder about time commitment. It does not mean any risk. Basically the data that is being collected is not at all delicate or confidential. It does not put the users' in any kind of harm. There is no data breach of any sort. All that is being collected is data regarding the time commitment that is not confidential at all and is being willingly provided by the users.

Hence, the exposure can be set at 100%

Duration = Total pageview/ Number of pageviews per day = 685325/40000 = 18 (approximately)

Experiment Analysis

Sanity Checks

For each of your 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. (These should be the answers from the "Sanity Checks" quiz.)

Sanity checks are conducted to check if the invariant metrics that are selected are reliable or not. These tests help establish that the invariant metrics are truly invariant that means they are truly divided equally among the experiment and the control group.

1. Number of cookies:

Total pageviews (Control Group) : 345543
Total pageviews (Experiment Group) : 344660
Total pageviews : 690203
Probability of cookie in control or experiment group: 0.5
SE = sqrt(0.5*(1-0.5)*(1/345543+1/344660) = 0.0006018
Margin of error = SE * 1.96 = 0.0011796
Confidence Interval = [0.5-m,0.5+m] = [0.4988,0.5012]
Observed value = 344660/690203 = 0.5006

2. Number of clicks:

Total Control group clicks : 28378
Total Experiment group clicks : 28325
Total pageview : 56703
Probability of cookie in control or experiment group : 0.5
SE = sqrt(0.5*(1-0.5)*(1/28378+1/28325) = 0.0021
Margin of error (m) = SE * 1.96 = 0.0041
Confidence Interval = [0.5-m,0.5+m] = [0.4959,0.5041]
Observed value = 28378/56703 = 0.50046

3. Click through probability:

CTP Control group: 28378 / 345543 = 0.082125813
CTP Exp group: 28325 / 344660 = 0.08218244
Pooled probability: 0.0822
SE pool: 0.0006610608156
Margin of error: 0.001295679199
Confidence Interval : -0.0013 to 0.0013
Observed Difference: 0.000056

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. Do not proceed to the rest of the analysis unless all sanity checks pass.

Comparing the confidence interval and the observed data in case of the three invariant metrics we can safely conclude that the sanity checks pass in all cases.

Result Analysis

Effect Size Tests

For each of your 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. (These should be the answers from the "Effect Size Tests" quiz.)

For every evaluation metric, we run effect size tests to check if the size of the effect is enough to make it statistically and practically important.

Gross conversion:
Confidence Interval: [-.0291, -.0120]
Dmin: 0.01
Statistically significant
Practically significant

Net conversion:
Confidence Interval: [-.0116, .0019]
Dmin: 0.0075
Not statistically significant
Not practically significant

Sign Tests

For each of your evaluation metrics, do 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. (These should be the answers from the "Sign Tests" quiz.)

Evaluation Metric	p-value	Significance
Gross conversion	***0.0026***	***Significant***
Net conversion	***0.6776***	***Insignificant***

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

I have not used Bonferroni correction. In statistics, the Bonferroni correction is one of several methods used to counteract the problem of multiple comparisons.

If we used just one metric out of the lot to make our launch decision, the Bonferroni method would be okay to use. But, because our hypothesis requires that 2 metrics be considered, Bonferroni correction would be too conservative.

There are no discrepancies between the effect size hypothesis and the sign tests.

Recommendation

Make a recommendation and briefly describe your reasoning.

Restating the hypothesis:

The **hypothesis** of adding the free trial screener was that this might set clearer expectations for students upfront,

reducing the number of frustrated students who left the free trial because they didn't have enough time → this can be interpreted as significant reduction in Gross Conversion

without significantly reducing the number of students to continue past the free trial and eventually complete the course → this can be interpreted as no significant reduction in Net Conversion.

The diagram below shows the evaluation metrics along with their mean value and the confidence interval. The dmin value is plotted for further clarity.

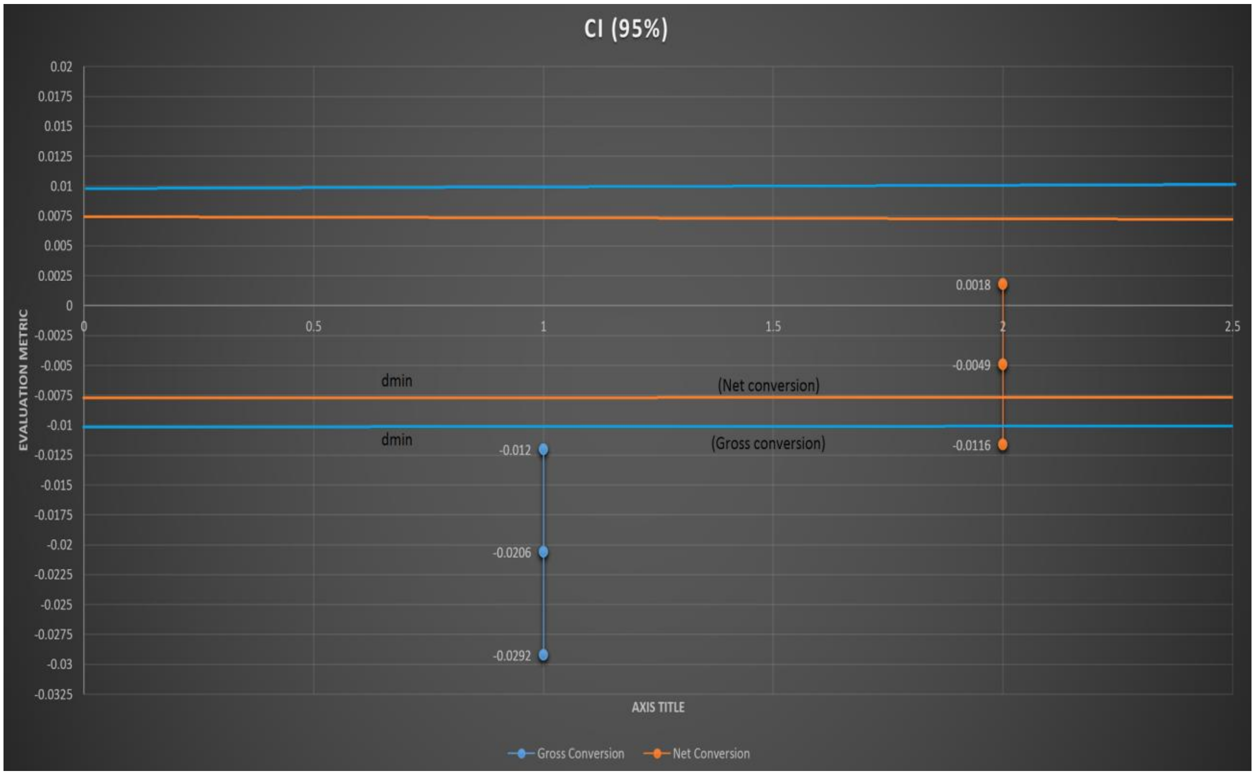
As the analysis indicates, change in Gross conversion turned out to be negative,statistically significant as well as and practically significant. This is a good outcome because we lower our costs by discouraging trial signups that are unlikely to convert.

Also, Net conversion is statistically and practically insignificant as demonstrated earlier. This means the number of students to continue past the free trial and eventually complete the course will not be significantly reduced.

However, the confidence interval for Net conversion deserves attention as it is important to look at the confidence interval for the net conversion rate to make sure that we aren't also introducing a negative effect with our manipulations.

CI for Net conversion is [-0.0116-0.0018]. Note that the lower interval in this case does cross the value for dmin indicating that even though the net conversion appears to be practically insignificant there is still some room for doubt. (There is some risk of reduction in revenue even though the estimates appear to be practically insignificant).

Hence, my recommendation would be to conduct additional tests.



Follow-Up Experiment

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.

The idea of using a free-trial screener is great. The information that the screener provides is about the time commitment. This assumes that the primary reason for the students leaving the course pre-payment and after starting the trail is that the time required cannot be given by them. However, in most cases this problem happens if the student finds the course too difficult or is unable to understand what is being taught because he/she lacks the understanding of prerequisites for the same. The trial period is spent in understanding the prerequisites and the student thereby ends up feeling like the course is too difficult for him/her.

This problem can be avoided if the screener provides the additional information regarding prerequisites, additional skills, and so on along with the time requirements. This would give the students enrolling a more complete understanding of what is to come. The course overview page does mention the pre-requisites. However, if a screener or prerequisite survey is taken before the student enrols it makes sure that the requirements are not overlooked. Also, if resources for learning the additional skills is provided that would contribute highly in reducing student’s frustration.

Hypothesis: The new screener with the information regarding the prerequisites and the skills needed to complete the course will reducing the number of frustrated students who left the free trial because they didn't have enough time and without significantly reducing the number of students to continue past the free trial and eventually complete the course.

Invariant Metrics: The invariant metric will be number of cookies and number of clicks as there will remain equal for both the groups.

Evaluation Metrics: The evaluation metric will be retention. A statistically and practically significant increase in Retention would indicate that the change is successful. This is in addition to Net conversion and Gross conversion. Significant reduction in Gross Conversion and no significant reduction in Net conversion is important for success.

Unit of Diversion: The unit of diversion will be cookies as using user id will not help us keep a tab on the users who did not complete the checkout but did see the screener.