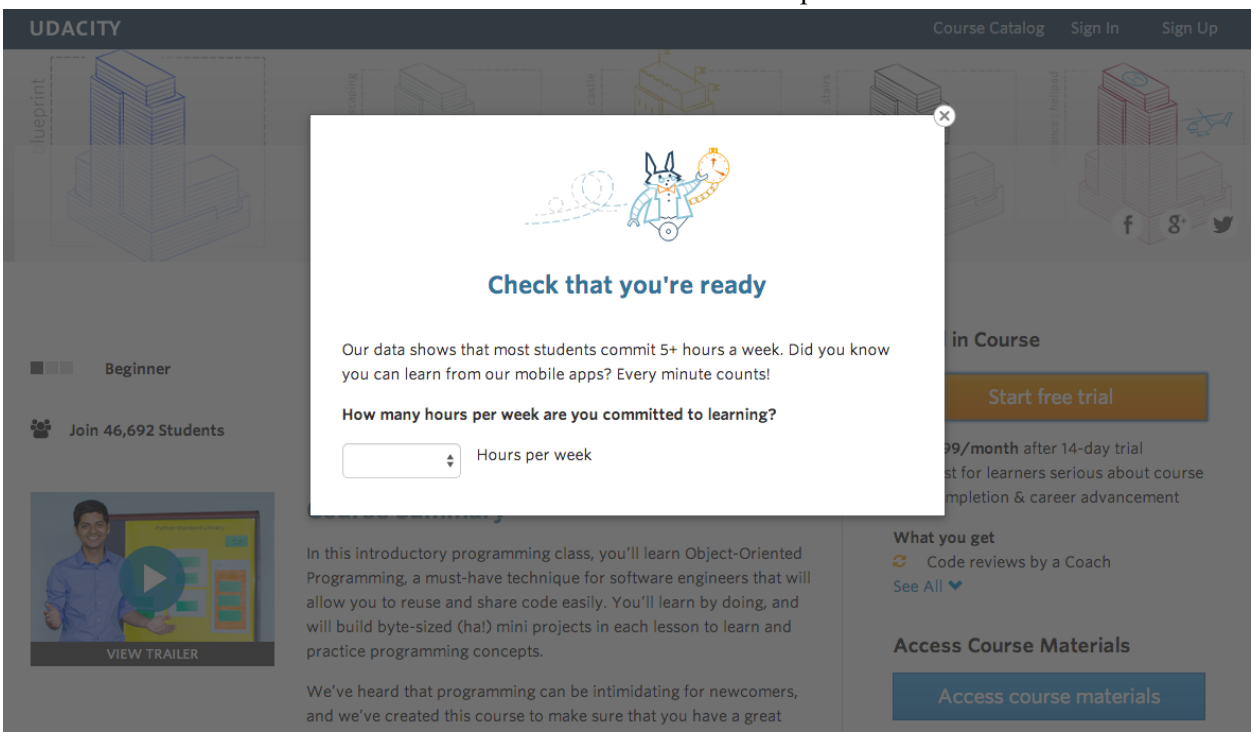
# Udacity A/B Testing Final Project

## Overview

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 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.

# Experiment Design

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

**Invariant metrics**

Invariant metrics are those metrics that remain invariant throughout the experiment and are expected to have similar distribution among control and experiment group.

(1) Number of cookies: That is, number of unique cookies to view the course overview page. This is the unit of diversion and should be evenly distributed among control and experiment group.
(2) 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). It is recorded before the experiment change occurred.
(3) 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.
Before the time the user clicks the "start free trial" button the user experience is same for all the users.

We did not choose number of user-ids as an invariant metric because user-ids are tracked only after enrolling in the free trial and the number of user-ids may be influenced by the experiment change.

**Evaluation metrics**

Evaluation metrics are the ones that we care about. Each evaluation metric is associated with a minimum difference ($d_{min}$) that must be observed for consideration in the decision to launch the experiment. This means it must be practically significant enough.

(1) 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. ($d_{min} = 0.01$)
(2) 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. ($d_{min} = 0.01$)
(3) 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. ($d_{min} = 0.0075$)

We expected the evaluation metrics as follow:

- Decreased gross conversion along with increase in net conversion, that is, less students enrolling in free trial but more students staying beyond the free trial.
- Increased retension, that is, the ratio of users who remained enrolled past the 14-day boundary to the number of users to complete checkout should increase.

## Measuring Standard Deviation

For each of the evaluation metrics, the standard deviation is calculated based on a sample size of 5000 unique cookies visiting the course overview page.
Baseline values for calculating standard deviation:

| | |
|---|---|
| Unique cookies to view course overview 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 |

$$\text{Standard Deviation of } \hat{P} = \sqrt{\frac{p(1-p)}{n}}$$

**(1) Gross conversion**

N=5000*0.08=400, p=0.20625
Standard deviation=0.0202

**(2) Retention**

N=5000*(660/40000)=82.5, p=0.53
Standard deviation=0.0549

**(3) Net Conversion**

N=5000*0.08=400, p=0.109313
Standard deviation=0.0156

## Sizing

### Number of Samples vs. Power

Because evaluation metrics are closely related to each other, Bonferroni would be too conservative to use in this case.
We used online sample size calculator (https://www.evanmiller.org/ab-testing/sample-size.html ) to calculate the number of samples needed. (alpha: 5%, beta: 20%)

**(1) Gross conversion**

Baseline rate: 20.625%, Minimum Detectable Effect: 1%
Sample size: 25,835 clicks/group, Total sample size: 25,835*2 = 51,670 clicks
Pageviews: 51,670 / 0.08 = 645,875

**(2) Retention**

Baseline rate: 53%, Minimum Detectable Effect: 1%
Sample size: 39,115 enrolls/group, Total sample size: 39,115*2 = 78,230 enrolls
Pageviews: 78,230 / (660 / 40,000) = 4,741,212

**(3) Net conversion**

Baseline rate: 10.93125%, Minimum Detectable Effect: 0.75%

Sample size: 27,413 clicks/group, Total sample size: 27,413*2 = 54,826 clicks
Pageviews: 54,826 / 0.08 = 685,325

The maximum number of pageviews required is 4,741,212.

### Duration vs. Exposure

Currently, udacity has 40000 pageviews per day. Even if we divert 100% traffic, the experiment will take about 119 days to run given the number of pageviews required for retention. Since it is too time-consuming, we can exclude retention and only use gross conversion and net conversion instead. This will take 18 days maximum if we use 100% of our traffic. Given that the duration is short, we can reduce the traffic for running the experiment to like 80% based on our need. This allowed our experiment to run at the same time.

# Experiment Analysis

The experiment data can be found in the link.

## Sanity Checks

To check whether the invariant metrics are equivalent between the two groups, we conducted sanity checks. We expected the number of cookies and clicks to be equally distributed among control group and experiment group.
We summarized the data as follow:

| | **Pageviews** | **Clicks** |
|---|---|---|
| **N_control** | 345543 | 28378 |
| **N_experiment** | 344660 | 28325 |
| **N_total** | 690203 | 56703 |
| **Probability (N_control/N_total)** | 0.5006 | 0.5005 |

Formula for calculation:
(1) The standard error for binomial distribution is:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

(2) Margin of error = 1.96 * SE (α = 0.05)
(3) Confidence interval bound = p ± margin

| | **Pageviews** | **Clicks** |
|---|---|---|
| **Observed probability** | 0.5006 | 0.5005 |
| **Standard error** | 0.0006 | 0.0021 |
| **Margin of error** | 0.0012 | 0.0041 |
| **CI lower bound** | 0.4988 | 0.4959 |
| **CI upper bound** | 0.5012 | 0.5041 |

We can see that the expected probability and observed probability for pageviews and clicks fall within the confidence interval, which means there is no significant difference between control group and experiment group. Therefore, these two metrics passed the sanity check.

For click-through-probability, we expect the difference between two group is zero. The formula for standard error calculation is:

$$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

| | **Click-through-probability** |
|---|---|
| **P_expected** | 0 |
| **P_control** | 0.082125814 |
| **P_experiment** | 0.082182441 |
| **P_observed** | 0.0000566 |
| **Standard error** | 0.00066 |
| **Margin of error** | 0.00129 |
| **CI lower bound** | -0.00129 |
| **CI upper bound** | 0.00129 |

The expected probability and observed probability for CTP also fall within the confidence interval. It passed the sanity check.

## Result Analysis

### Effect Size Tests

For each evaluation metrics, we will calculate the confidence interval for the difference between the control and experiment groups based on 95% confidence interval and check whether each metric is statistically or practically significant.
A metric is statistically significant if the confidence interval does not include 0 and it is practically significant if the confidence interval does not include the practical significance.
From the previous analysis, we chose the Gross Conversion and Net Conversion as our final evaluation metrics.
Data summary as follow:

| | **Control** | **Experiment** |
|---|---|---|
| **Clicks** | 17293 | 17260 |
| **Enrollments** | 3785 | 3423 |
| **Payment** | 2033 | 1945 |

| | **Gross conversion** |
|---|---|
| **P_control** | 0.21887469 |
| **P_experiment** | 0.19831981 |
| **P_pooled** | 0.20860707 |
| **Standard error_pooled** | 0.00437168 |
| **Margin of error_pooled** | 0.00856848 |
| **dmin** | 0.01 |
| **Observed difference** | -0.02055488 |
| **CI lower bound** | -0.02912336 |

| | |
|---|---|
| **CI upper bound** | -0.01198640 |

Because the confidence interval does not include zero and dmin, the gross conversion is both statistically significant and practically significant.

| | **Net conversion** |
|---|---|
| **P_control** | 0.11756202 |
| **P_experiment** | 0.11268830 |
| **P_pooled** | 0.11512749 |
| **Standard error_pooled** | 0.00343413 |
| **Margin of error_pooled** | 0.00673090 |
| **dmin** | 0.0075 |
| **Observed difference** | -0.00487372 |
| **CI lower bound** | -0.01160462 |
| **CI upper bound** | 0.00185718 |

Because the confidence interval includes zero and negative dmin, the gross conversion is neither statistically significant nor practically significant.

### Sign Tests

We performed sign test using day-by-day data. The sign test 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.
We used online calculator to get the result. (https://www.graphpad.com/quickcalcs/binomial1/)
For Gross Conversion, there are 4 days during which enrollments in experiment group are higher than those in control group. p-value = 0.0026, which is significant.
For Net Conversion, there are 10 days during which enrollments in experiment group are higher than those in control group. p-value = 0.6776, which is insignificant.

### Summary

We conducted an experiment to test whether asking the amount of time the students available to devote to study after clicking a “start free trail button” will have positive impact, this is, improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.
Three metrics (Number of Cookies, Number of clicks on "start free trial", and Click-Through-Probability) were selected as invariant metrics for validation and passed the sanity check. Gross Conversion and Net Conversion were served as evaluation metrics.
The null hypothesis is that there is no difference in the evaluation metrics between the control group and experiment group. Furthermore, a practical significance threshold was set for each evaluation metric.
The requirement for launching the experiment is that the null hypothesis must be rejected for all evaluation metrics and that the difference between two groups must meet or exceed the practical significance threshold using 95% confidence interval. Under this circumstance, the use of Bonferonni correction is not appropriate because it is a method for controlling for type I errors (false positives) when using multiple metrics in which relevance of any of the metrics matches the hypothesis.
Analysis indicated that the gross conversion is both statistically significant and practically significant while the gross conversion is neither statistically significant nor practically significant.

## Recommendation

This experiment was designed to determine whether filtering out students who did not have enough study time commitment will improve the overall student experience and the coaches' capacity to support students who are likely to complete the course, without significantly reducing the number of students who will make the payment after completing their free trial. Our analysis indicated that the gross conversion will be reduced significantly but there is no significant change in Net Conversion. Therefore, the filter will only help reduce the enrollment, but will not increase the number of students who make the payments. Given this, we suggested not to launch the experiment.

# Follow-Up Experiment

We define the frustrated student who cancel early in the course as the student who cancel prior to the end of the 14-day trial period in which payment is triggered. Cancellation rate can be defined as gross conversion minus net conversion. The reasons for early cancellation can be summarized to two categories: (1) the students do not have enough time to devote to study so that they tend to give up during trial period (2) the course itself may not meet the students' needs or expectations. The students have issues but cannot find any help.

For the first category, we can conduct similar experiment as we did before: filtering out those students to see whether the cancellation rate will decrease. The process is almost the same so we will focus on experiment design for the second category.

For the second category, changing the course content is not practical in short-term so we can test whether providing coach will reduce the cancellation rate.

**Experiment**

Whether providing coach service will contribute to the decrease of cancellation rate. For students that are stuck in the quizzes, in experiment group, we set a trigger system to suggest students to use coach service while in control group, we do not pop up any message.

**Hypothesis**

The hypothesis is that the trigger system will help frustrated student complete the course. If the hypothesis is true, the experiment will have a higher Net conversion. Therefore, we have a lower Cancellation rate.

**Unit of diversion**

cookies

**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)

- 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)