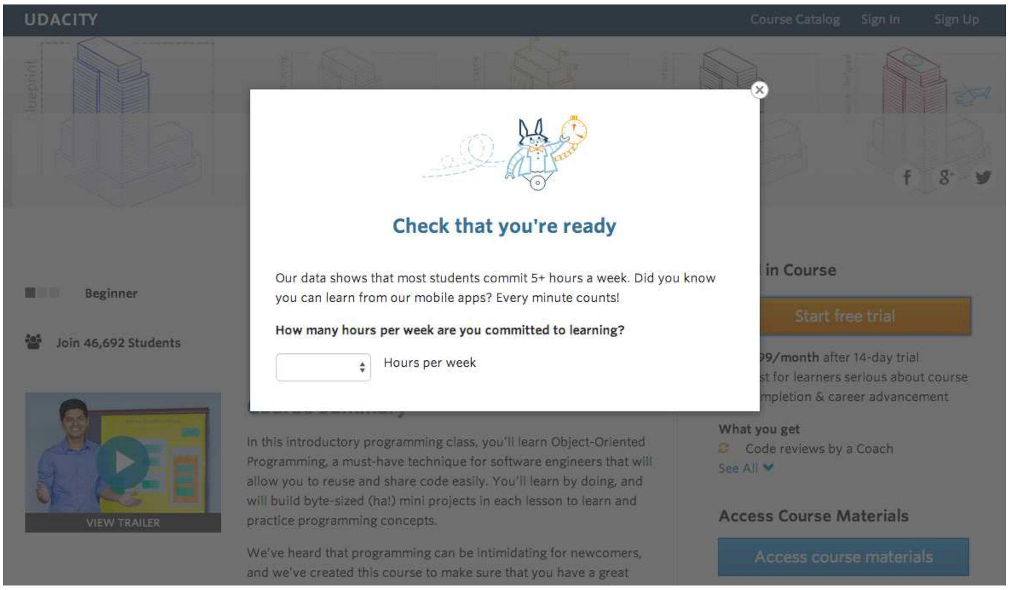
PROJECT: Design an A/B Test – Free Trial Screener

Experiment 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. Figure 1 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

99/month after 14-day trial
st for learners serious about course
mpletion & career advancement

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

FIGURE 1 – Free-Trial screener.

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

Table 1 contains a list of available metrics that can be measured for this experiment. 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.

Metric	**Description**	**Practical Significance Boundary (d_{min})**
Number of cookies	The number of unique cookies to view the course overview page.	3000
Number of user-ids	The number of users who enroll in the free trial.	50
Number of clicks	The number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger).	240
Click-through-probability	The 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
Gross conversion	The 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
Retention	The 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.	0.01
Net conversion	The 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.	0.0075

TABLE 2 – Available metrics.

Invariant Metrics

Invariant metrics are metrics that are expected to be unaffected by the experiment, resulting in an even distribution between the control and experiment groups. If a change is observed in an invariant metric then there must me something wrong with the experiment.

- **Number of cookies:** This is the unit of diversion. This metric is independent of the experiment since the visit to the course overview page occurs before the free trial screener is trigger.

- **Number of clicks:** This metric is also independent of the experiment since the free trial screener is triggered after the "Start free trial" button is clicked.

- **Click-through-probability:** This metric is the ratio of the above two metrics. Since they are both independent and occur before the free trial screener is trigger, this metric is also independent.

Evaluation Metrics

Evaluation metrics are metrics that we expect to be affected by the experiment, resulting in different distributions between the control and experiment groups.

- **Gross conversion:** This metric is a measure of enrollments. The goal of the experiment is to dissuade users who are likely to drop out during the free trial period from enrolling and possibly consuming limited coaching resources.

- **Net conversion:** This metric is a measure of payments. I addition to the above goal, the experiment should not have a significant impact on the number of users to remain enrolled after the free trial period (and thus make at least one payment).

Unused Metrics

- **Number of user-ids:** The number of users who enroll in the free trial is expected to be different between the control and experiment groups, but is not normalised. We are already measuring this by using gross conversion as an evaluation metric since it is normalised with the unit of diversion.

- **Retention:** This is simply the ratio of net conversion to gross conversion. We will not use this metric since we are already using both of those.

In order to launch the experiment, we would want to see a decrease in the gross conversion metric by a significant amount (both statistically and practically), without seeing any significant change in the net conversion metric.

Measuring Variability

Table 2 contains rough estimates of the baseline values for these metrics.

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

TABLE 2 - Baseline values.

For each evaluation metric, we can make an analytic estimate of its variability given a sample size of 5000 cookies visiting the course overview page. Adjusting for the sample size:

$$N = 5000 \times 0.08 = 400$$

For the sample, assuming a binomial distribution:

$$SE = \sqrt{\frac{\hat{P} \times \left(1 - \hat{P}\right)}{N}}$$

$$SE_{GROSS\ CONVERSION} = \sqrt{\frac{0.20625 \times (1 - 0.20625)}{400}} = \mathbf{0.0202}$$

$$SE_{NET\ CONVERSION} = \sqrt{\frac{0.1093125 \times (1 - 0.1093125)}{400}} = \mathbf{0.0156}$$

We would expect the analytic estimates to be close to the empirical estimates of the variability because the unit of analysis is the same as the unit of diversion for each metric.

Sizing

Number of Samples given Power

Using the analytic estimates of variance, the total number of pageviews (across both groups, experiment and control) needed to adequately power the experiment can be estimated using the following online calculator: http://www.evanmiller.org/ab-testing/sample-size.html.

Gross Conversion:

$$\alpha = 0.05, \qquad \beta = 0.20$$

$$Baseline\ Conversion\ Rate = 0.20625$$

$$Minimum\ Detectable\ Effect = d_{min} = 0.01$$

From: http://www.evanmiller.org/ab-testing/sample-size.html#!20.625;80;5;1;0

$$Sample\ Size\ (per\ group) = 25835$$

$$Total\ Sample\ Size = 2 \times 25835 = 51670$$

$$Pageviews = \frac{51670}{0.08} = \mathbf{645875}$$

Net Conversion:

$$\alpha = 0.05, \qquad \beta = 0.20$$

$$Baseline\ Conversion\ Rate = 0.1093125$$

$$Minimum\ Detectable\ Effect = d_{min} = 0.0075$$

From: http://www.evanmiller.org/ab-testing/sample-size.html#!10.93125;80;5;0.75;0

$$Sample\ Size\ (per\ group) = 27413$$

$$Total\ Sample\ Size = 2 \times 27413 = 54826$$

$$Pageviews = \frac{54826}{0.08} = \mathbf{685325}$$

To ensure we have enough power for each metric we use the largest of these estimates.

Minimum number of pageviews required: **685325**

Duration vs. Exposure

To minimise the duration of the experiment I suggest diverting 100% of Udacity's traffic to this experiment (assuming there were no other experiments to be run simultaneously). The change constitutes no more than minimal risk, and does not deal with sensitive data.

Given 40000 pageviews/day we can calculate the expected duration of the experiment using the minimum number of pageviews required:

$$\frac{685325}{40000} = 17.1331 = \boldsymbol{18\ days}\ (\boldsymbol{rounded\ to\ whole\ days})$$

This is reasonable since it only extends slightly longer than the free trial period.

Analysis

The data for this experiment is available at:

https://docs.google.com/a/knowlabs.com/spreadsheets/d/1Mu5u9GrybDdska-ljPXyBjTpdZIUev_6i7t4LRDfXM8/edit#gid=0

This data contains the raw information needed to compute the above metrics, broken down day by day. There are two sheets within the spreadsheet - one for the experiment group, and one for the control group. Table 3 details the meaning of each column.

Column	Description
Pageviews	Number of unique cookies to view the course overview page that day.
Clicks	Number of unique cookies to click the course overview page that day.
Enrollments	Number of user-ids to enroll in the free trial that day.
Payments	Number of user-ids who enrolled on that day to remain enrolled for 14 days and thus make a payment. (Note that the date for this column is the start date, that is, the date of enrollment, rather than the date of the payment. The payment happened 14 days later. Because of this, the enrollments and payments are tracked for 14 fewer days than the other columns).

TABLE 3 – Description of data.

Sanity Checks

As a sanity check we will check whether the invariant metrics are equivalent between the control and experiment groups. From the data:

$$X_C = \sum(Control\ Pageviews) = 28378$$

$$N_C = \sum(Control\ Clicks) = 345543$$

$$X_E = \sum(Experiment\ Pageviews) = 28325$$

$$N_E = \sum(Experiment\ Clicks) = 344660$$

The number of cookies and number of clicks are both simple counts that should be randomly split between the control and experiment groups. For these we will use a binomial test.

Number of Cookies:

$$\hat{P}_C = \frac{N_C}{N_C + N_E} = \frac{345543}{345543 + 344660} = \mathbf{0.5006}$$

$$P = 0.5$$

For a 95% confidence interval,

$$\alpha = 0.05, \qquad Z^*_{\alpha/2} = 1.96$$

$$SE = \sqrt{\frac{P \times (1 - P)}{X_C + X_E}} = \sqrt{\frac{0.5 \times (1 - 0.5)}{345543 + 344660}} = 0.0006$$

$$m = Z^*_{\alpha/2} \times SE = 1.96 \times 0.0006 = 0.0012$$

$$Lower\ Bound = P - m = 0.5 - 0.0012 = \mathbf{0.4988}$$
$$Upper\ Bound = P + m = 0.5 + 0.0012 = \mathbf{0.5012}$$

This metric passes the sanity check since the observed fraction of 0.5006 is between the bounds of the confidence interval: $0.4988 < \hat{P}_C < 0.5012$.

Number of Clicks:

$$\hat{P}_C = \frac{X_C}{X_C + X_E} = \frac{28378}{28378 + 28325} = \mathbf{0.5005}$$

$$P = 0.5$$

For a 95% confidence interval,

$$\alpha = 0.05, \qquad Z^*_{\alpha/2} = 1.96$$

$$SE = \sqrt{\frac{P \times (1-P)}{X_C + X_E}} = \sqrt{\frac{0.5 \times (1-0.5)}{28378 + 28325}} = 0.0021$$

$$m = Z^*_{\alpha/2} \times SE = 1.96 \times 0.0021 = 0.0041$$

$$Lower\ Bound = P - m = 0.5 - 0.0041 = \mathbf{0.4959}$$
$$Upper\ Bound = P + m = 0.5 + 0.0041 = \mathbf{0.5041}$$

This metric passes the sanity check since the observed fraction of 0.5005 is between the bounds of the confidence interval: $0.4959 < \hat{P}_C < 0.5041$.

For the click through probability we will need to construct a confidence interval for a difference in proportions, then check whether the difference between group values falls within that confidence level.

Click Through Probability:

$$\hat{P}_C = \frac{X_C}{N_C} = \frac{28378}{345543} = 0.0821$$

$$\hat{P}_E = \frac{X_E}{N_E} = \frac{28325}{344660} = 0.0822$$

$$\hat{P}_{POOL} = \frac{X_C + X_E}{N_C + N_E} = \frac{28378 + 28325}{345543 + 344660} = 0.0822$$

$$SE_{POOL} = \sqrt{\hat{P}_{POOL} \times \left(1 - \hat{P}_{POOL}\right) \times \left(\frac{1}{N_C} + \frac{1}{N_E}\right)}$$
$$= \sqrt{0.0822 \times (1 - 0.0822) \times \left(\frac{1}{345543} + \frac{1}{344660}\right)} = 0.0007$$

$$\hat{d} = \widehat{P_E} - \widehat{P_C} = 0.0822 - 0.0821 = \mathbf{0.0001}$$

For a 95% confidence interval,

$$\alpha = 0.05, \qquad Z^*_{\alpha/2} = 1.96$$

$$m = Z^*_{\alpha/2} \times SE_{POOL} = 1.96 \times 0.0007 = 0.0013$$

$$Lower\ Bound = 0 - m = 0 - 0.0013 = \mathbf{-0.0013}$$
$$Upper\ Bound = 0 + m = 0 + 0.0013 = \mathbf{0.0013}$$

This metric passes the sanity check since the difference of 0.0001 is between the bounds of the confidence interval: $-0.0013 < \hat{d} < 0.0013$.

Practical and Statistical Significance

For each evaluation metric, we can calculate the confidence interval for the difference between the experiment and control groups, to check whether it is statistically and/or practically significant.

Gross Conversion:

From the data:

$$X_C = \sum (Control\ Enrollments) = 3785$$

$$N_C = \sum (Control\ Clicks) = 17293$$

$$X_E = \sum (Experiment\ Enrollments) = 3423$$

$$N_E = \sum (Experiment\ Clicks) = 17260$$

$$\hat{P}_C = \frac{X_C}{N_C} = \frac{3785}{17293} = 0.2189$$

$$\hat{P}_E = \frac{X_E}{N_E} = \frac{3423}{17260} = 0.1983$$

$$\hat{P}_{POOL} = \frac{X_C + X_E}{N_C + N_E} = \frac{3785 + 3423}{17293 + 17260} = 0.2086$$

$$SE_{POOL} = \sqrt{\hat{P}_{POOL} \times (1 - \hat{P}_{POOL}) \times \left(\frac{1}{N_C} + \frac{1}{N_E}\right)}$$

$$= \sqrt{0.2086 \times (1 - 0.2086) \times \left(\frac{1}{17293} + \frac{1}{17260}\right)} = 0.0044$$

$$\hat{d} = \widehat{P_E} - \widehat{P_C} = 0.1983 - 0.2189 = -0.0206$$

For a 95% confidence interval,

$$\alpha = 0.05, \qquad Z^*_{\alpha/2} = 1.96$$

$$m = Z^*_{\alpha/2} \times SE_{POOL} = 1.96 \times 0.0044 \ = 0.0086$$

$$Lower\ Bound = \hat{d} - m = -0.0206 - 0.0086 = \mathbf{-0.0291}$$
$$Upper\ Bound = \hat{d} + m = -0.0206 + 0.0086 = \mathbf{-0.0120}$$

$$d_{min} = \pm 0.01$$

Gross conversion is statistically significant since the confidence interval does not include 0 (that is, I am confident there was a change). It is also practically significant since the confidence interval does not include the practical significance boundary, d_{min} (that is, I am confident there is a change that matters to the business).

Net Conversion:

From the data:

$$X_C = \sum(Control\ Payments) = 2033$$

$$N_C = \sum(Control\ Clicks) = 17293$$

$$X_E = \sum(Experiment\ Payments) = 1945$$

$$N_E = \sum(Experiment\ Clicks) = 17260$$

$$\hat{P}_C = \frac{X_C}{N_C} = \frac{2033}{17293} = 0.1176$$

$$\hat{P}_E = \frac{X_E}{N_E} = \frac{1945}{17260} = 0.1127$$

$$\hat{P}_{POOL} = \frac{X_C + X_E}{N_C + N_E} = \frac{2033 + 1945}{17293 + 17260} = 0.1151$$

$$SE_{POOL} = \sqrt{\hat{P}_{POOL} \times (1 - \hat{P}_{POOL}) \times \left(\frac{1}{N_C} + \frac{1}{N_E}\right)}$$

$$= \sqrt{0.1151 \times (1 - 0.1151) \times \left(\frac{1}{17293} + \frac{1}{17260}\right)} = 0.0034$$

$$\hat{d} = \widehat{P_E} - \widehat{P_C} = 0.1127 - 0.1176 = -0.0049$$

For a 95% confidence interval,

$$\alpha = 0.05, \qquad Z^*_{\alpha/2} = 1.96$$

$$m = Z^*_{\alpha/2} \times SE_{POOL} = 1.96 \times 0.0034 \ = 0.0067$$

$$Lower\ Bound = \hat{d} - m = -0.0049 \ - 0.0067 = \mathbf{-0.0116}$$
$$Upper\ Bound = \hat{d} + m = -0.0049 \ + 0.0067 = \mathbf{0.0019}$$

$$d_{min} = \pm 0.0075$$

Net conversion is not statistically significant since the confidence interval does include 0 (that is, I am not confident there was a change). It is also not practically significant since the confidence interval does include the negative practical significance boundary, d_{min} (that is, I am not confident there is a change that matters to the business).

The Bonferroni correction was not used in this experiment. It would be too conservative since the metrics are highly correlated.

Sign Tests

For each evaluation metric, we will carry out a sign test using the day-by-day breakdown. For each day of the experiment we will calculate the difference in each metric between the control and experiment groups.

Assuming a binomial distribution since there are two possible outcomes, and assuming a probability of 0.5 since there is an equal chance of a positive or negative change on each day, we will calculate the two-tailed P-value for the metric using the online calculator:

https://www.graphpad.com/quickcalcs/binomial1/

Gross Conversion:

The goal of the experiment is to see a reduction in this metric. From the data:

Number of "successes" (negative change): 19
Number of trials per experiment: 23

Sign test – If the probability of "success" in each trial is 0.500, then:

The one-tail P-value is **0.0013**
This is the chance of observing 19 or more successes in 23 trials.

The two-tail P-value is **0.0026**
This is the chance of observing either 19 or more successes, or 4 or fewer successes, in 23 trials.

The change **is** statistically significant since the two-tail P-value is smaller than $\alpha = 0.05$.

Net Conversion:

The goal of the experiment is also to minimise any reduction in this metric. From the data:

Number of "successes" (positive change): 10
Number of trials per experiment: 23

Sign test – If the probability of "success" in each trial is 0.500, then:

The one-tail P value is **0.3388**
This is the chance of observing 10 or fewer successes in 23 trials.

The two-tail P value is **0.6776**
This is the chance of observing either 10 or fewer successes, or 13 or more successes, in 23 trials.

The change **is not** statistically significant since the two-tail P-value is larger than $\alpha = 0.05$.

Recommendation

In order to launch the experiment, we would want to see a decrease in the gross conversion metric by a significant amount (both statistically and practically), without seeing any significant change in the net conversion metric. The results of our tests confirm that this was observed for both metrics. However, for net conversion, the effect size test showed that the 95% confidence interval includes the negative practical significance boundary, meaning there is a possibility of this metric presenting a practically significant decrease. My recommendation therefore would be **not to launch** this experiment.

Follow-Up Experiment: How to Reduce Early Cancellations

If we want to reduce the number of frustrated students who cancel early in the course, we must first try to understand the reasons behind this. Some students might cancel early because the course failed to live up to their expectations. It may have been their first experience of an online course and didn't suit them. They may have been lacking in pre-requisite skills needed to complete the course in the required time-frame. It may even have been the uncertainty of return on their investment, i.e. whether successfully completing the course would enhance their career.

Not all students learn the same way or at the same pace, so the 14-day free trial may not be enough for them to start to make reasonable progress and begin to see the benefits. My suggestion for a follow-up experiment would be to change the structure of the introductory period from a 14-day free trial to a free 1st introductory course in a larger program like the nanodegree. This would allow students who naturally take longer to attain the same tuition at their own pace. The benefit of this would be that by the time the students are paying for the 2nd course in the program, they should be more familiar with the course structure and have comparable pre-requisite skills. They would be more engaged and more likely to complete the entire program.

For this experiment, the null hypothesis would be that the change had no effect on the number of students to continue onto the paid-for section of the program. The alternate hypothesis would be that the change significantly increased the number of students to progress through to the paid-for section of the program.

To measure the effect of the experiment I would use the following evaluation metrics:

Completion – The number of user-ids to successfully complete the free course divided by the number of user-ids to enroll in the free course.

Conversion – The number of user-ids to enroll on the 2nd course in the program (and thus make at least one payment) divided by the number of user-ids to enroll in the free course.

Since the student would have to be logged in to enroll on a course, the unit of diversion would be **number of user-ids**, in other words the number of users who enroll in the program.

References

http://www.evanmiller.org/ab-testing/sample-size.html

https://www.graphpad.com/quickcalcs/binomial1/