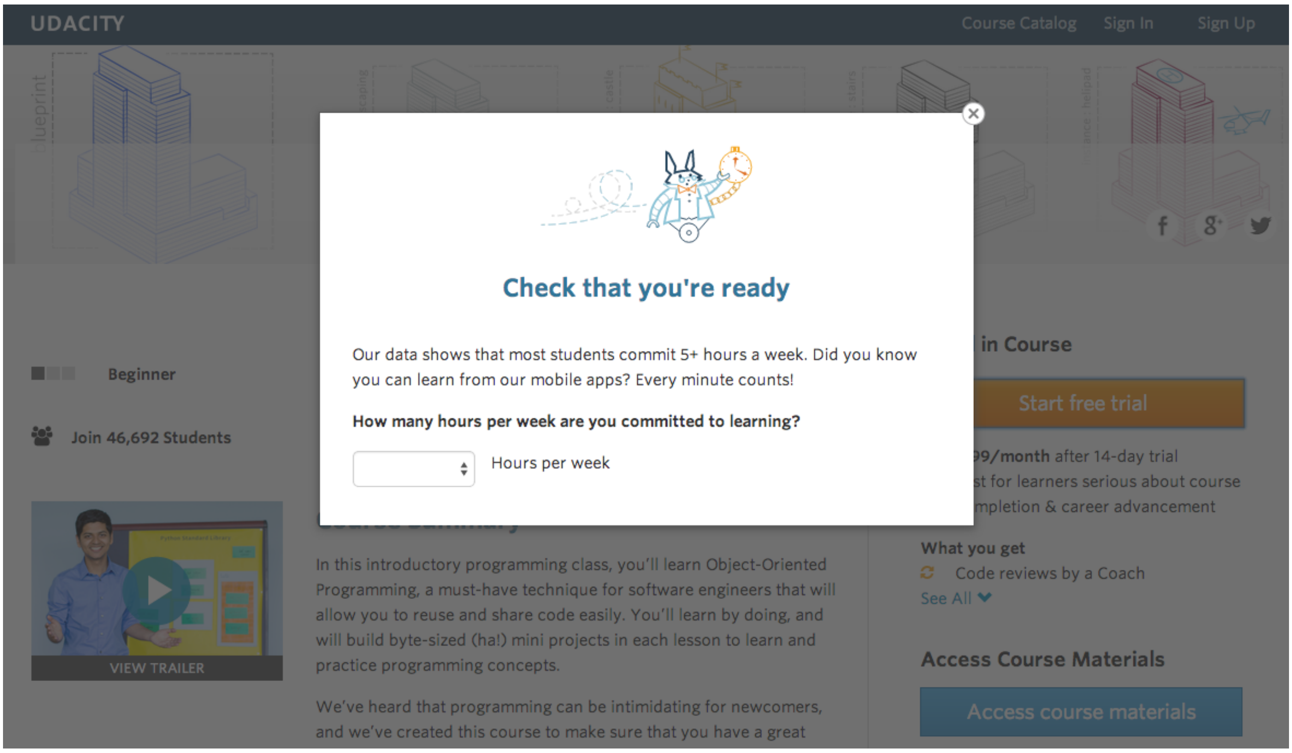
A/B Test Project

Experiment 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 (see screenshot below).



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

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

The seven metrics provided as choice for the project are as follows:

- **Number of cookies:** That is, number of unique cookies to view the course overview page. (dmin=3000)
- **Number of user-ids:** That is, number of users who enroll in the free trial. (dmin=50)
- **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)

Used Metrics

As far as the invariant metrics are concerned, I chose the “Number of cookies”, the “Number of clicks”, and the “Click-through probability” because they are all metrics that measure activity before the differences between the control and experiment groups come into effect. In addition, they are all based on a measurement of cookies, which is also the unit of diversion.

For the evaluation metrics, “Gross conversion” and “Net conversion” are the best metrics. As is stated in the project overview, the point of the experiment is to “...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**.”

Both metrics measure the Number of clicks, with Gross conversion measuring with how many people actually sign up and start the free trail compared to the Number of clicks, and Net conversion measuring how many people actually finish the free trail compared to the Number of clicks. Therefore, both metrics directly measure what the experiment is trying to achieve. In order for these metrics to pass my analysis and allow me to recommend a launch. They need to meet the following criteria:

- **Gross conversion:** in order for this metric to pass my analysis it needs to prove to have been both Statistically and Practically significant in the negative direction. This is due to the experiments hypothesis that the message barrier about time commitment will deter people who are not ready to take the courses from signing up, so we expect less people to sign up in the experiment group compared to the control group.
- **Net conversion:** in order for this metric to pass my analysis it needs to not be practically significant in the negative direction. This is due to the hypothesis stating that the message barrier would not “significantly reducing the number of students to continue past the free trial and eventually complete the course.”

Unused Metrics

I did not use “Number of user-ids” as an invariant metric because it will be generated after the experimental message warning visitors about the recommended learning time is shown. As a result, it is likely to differ significantly between the control and experiment groups. Because of this, it is possible this could have been used as an evaluation metric, but the probabilities that use this metric (such as Gross conversion, Retention, and Net conversion) give greater insight than raw numbers alone.

I did not use “Retention” as an evaluation metric because of the amount of time it would take to use it. Later in the “Sizing” section I note how it would take around 120 days to get enough data to use this metric, compared to 18 days for Gross and Net retention. As a result, it is not feasible to use this metric within a reasonable time-frame.

Measuring Variability

First, I need to consider if I need to calculate the analytical and empirical estimates of the standard deviations (SD) for each metric, and I have decided that calculating only the analytical estimates is fine. This is due to the nature of the underlying data and the metrics I have chosen. The data itself is assumed to be binomial because my metrics are probabilities (which have normal distributions with large enough samples), and the unit of my analysis for both evaluation metrics is the same as the unit of diversion. In both of these cases, calculating the empirical estimates is not necessary.

In order get analytical estimates of the SD for my chosen evaluation metrics, I used the baseline values shown below:

Unique cookies to view course overview page per day (Number of cookies):	40000
Unique cookies to click "Start free trial" per day (Number of clicks):	3200
Enrollments per day (Number of user-ids):	660
Click-through-probability on "Start free trial":	0.08
Probability of enrolling, given click (Gross conversion):	0.20625
Probability of payment, given enroll (Retention):	0.53
Probability of payment, given click (Net Conversion):	0.1093125

I then recalculated these numbers using a sample size of 5000 unique cookies to view the course overview page, giving me the following numbers for the first three values:

Unique cookies to view course overview page per day (Number of cookies):	5000
Unique cookies to click "Start free trial" per day (Number of clicks):	400
Enrollments per day (Number of user-ids):	82.5

Given this formula for finding an SD,

$$SD=\sqrt{\frac{p*(1-p)}{N}}$$

the results are as follows:

Gross conversion	0.0202
Net conversion	0.0156

Sizing

Choosing Number of Samples given Power

In order to calculate how many total page-views I will need to get enough data for statistically powerful results, I used this online calculator. Given a baseline conversion rate of 10.93125%, a minimum detectable effect of 0.75%, an alpha (α) level of 0.5, and a beta (β) level of 0.2, the number of total clicks needed on the “Start free trial” button came out to 27,413 clicks.

Given a click-through rate of 0.08 and clicks at 27,413, the total page-views needed for the experiment were calculated like so:

$$Pageviews = \left(\frac{27{,}413}{0.08}\right) * 2 = 685{,}325$$

Choosing Duration vs. Exposure

In order to figure out how long to run the study, I looked to the baseline value of 40,000 for page-views and the total 685,325 page-views needed for the experiment. The length of the experiment will depend on what percentage of the baseline page-views I want to direct into the experiment. Assuming that there are no other experiments I would want to run simultaneously with this one, the quickest approach would be to direct all of the 40,000 page-views into the experiment, splitting the page-views evenly into the control and experiment groups. If we do that, the length of the experiment would be equal to just over 17 days given by the calculation shown below:

$$Length\ of\ experiment = \frac{40{,}000}{685{,}325} = 17.133125$$

As a result, the experiment should be run for a full **18 days** in order to achieve the desired total page-views.
(**Note:** I had considered using Retention has an an evaluation metric a well, but it was at this point that I dropped it. During my sizing calculations I discovered that it would have taken about 120 days of diverting the full 40,000 page-views to generate the enrollments needed to make the metric viable.)

In terms of website functionality, running the experiment on all the traffic is risky in principal. However, for this particular experiment I think it will be fine. Only half of the traffic would actually be experiencing any change, and the change is relatively minor to the overall functioning of the website. In terms of the ethics of the experiment, I think there is little risk to any individual user, so having a large number of users exposed to the experiment would not pose any ethical issues. The data that is collected, namely whether a user would have time to devote 5 hours or more per week to a course, is not particularly sensitive information.

Analysis

Sanity Checks

Now that the experiment has been run, it’s time to analyze the results. First I will look at each invariant metric, calculate the confidence interval (CI) to get a range of the values I expect to observe, and calculate the actual observed value to determine if it falls within the expected range. Below is a chart of some of the relevant data needed for the calculations.

	Page-views	**Clicks**
Control	345543	28378
Experiment	344660	28325
Control + Experiment	690203	56703

In order to calculate the CI for total page-views and clicks, I first have to find the SD for each value. The formula for finding the SD is the same as before, and the expected probability for the totals in both the Control and Experiment groups is 0.5:

$$Pageviews\ SD = \sqrt{\frac{0.5*(1-0.5)}{690{,}203}} \qquad Clicks\ SD = \sqrt{\frac{0.5*(1-0.5)}{5{,}6703}}$$

Page-views SD	0.00060184
Clicks SD	0.00209975

I then find each margin of error (ME) by multiplying each SD by the Z-score for a two-tailed test with an α level of 0.05:

$$ME = SD * 1.96$$

Page-views ME	0.00117961
Clicks ME	0.00411550

For the confidence interval we take the 0.5 probability mentioned above and add and subtract the ME for both page-views and clicks:

$$Lower\ \ Bound = 0.5 - \ ME \qquad Upper\ \ Bound = 0.5 + \ ME$$

	Lower Bound	**Upper Bound**
Page-views	0.4988	0.5012
Clicks	0.4959	0.5041

Finally, I will calculate the actual observed page-view and click probability by dividing the control over the the total and determine a pass or fail for each observed probability if they fall withing the range of the CI :

$$Pageviews\ \ Observed = \frac{345543}{690203} \qquad Clicks\ Observed = \frac{28378}{56703}$$

	Observed	**Pass/Fail**
Page-views	0.5006	Pass
Clicks	0.5004	Pass

The procedure for calculating the SD for the click-through probability, however, is a little different. In fact, the SD for probabilities is called the Standard Error (SE). First I need to calculate the

Pooled Click-through Probability (Ppool), and then use that to calculate the Pooled Standard Error (SEpool):

$$P\hat{p}ool = \frac{56{,}703}{690{,}203} = 0.0822$$

$$SEpool = \sqrt{0.0822 * (1 - 0.0822) * (\frac{1}{345543} + \frac{1}{344660})} = 0.0007$$

Next, I will find the CI by multiplying the SEpool by 1.96 and -1.96:

$$Lower\ \ Bound = 0.0007 * -1.96 \qquad Upper\ \ Bound = 0.0007 * 1.96$$

	Lower Bound	Upper Bound
Page-views	-0.0013	0.0013

Finally, I need to find the difference ($\hat{d}$) between the click-through probability for both the Control ($\hat{P}$ cont) and Experiment ($\hat{P}$ exp) groups:

$$\hat{d} = 0.00001$$

$\hat{d}$ is within the range of the CI.

Check for Practical and Statistical Significance

Now it is time to check if my evaluation metrics of Gross conversion and Net conversion are statistically and practically significant. A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident 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 there is a change that matters to the business).

In order to do this, I will perform the same calculations I did above for the click-through probability. However, the numbers for clicks had to be recalculated. This is due to how the enrollments and payments were tracked. Because payment happened 14 days after the date of enrollment, the enrollments and payments are tracked for 14 fewer days than the other columns. As a result, clicks for the control group and experiment groups went down:

	Clicks
Control	28378
Experiment	28325
Control + Experiment	56703

→

Clicks
17293
17260
34553

The calculations for $\hat{P}$ pool and SEpool were the same as the calculations for the Click-through Probability shown above, but differed on how to find the CI. Instead of simply multiplying the SEpool by 1.96 and -1.96, I first had to find the ME and then add and subtract the ME with $\hat{d}$:

	Lower Bound	Upper Bound	dmin	Statistically Significant?	Practically Significant?
Gross conversion	-0.0291	-0.0119	0.01	Yes	Yes
Net conversion	-0.0116	0.0018	0.0075	No	No

As shown above, the CI for Gross conversion was both statistically and practically significant, but the CI for Net conversion was not.

Run Sign Tests

Next I will run a sign test. In order to do that I will calculate the daily Gross and Net conversion rates for the control and experiment groups. For Gross conversion I will tally up the number of successful days where the Control Gross conversion is greater than the Experiment Gross conversion, and for Net conversion it will be the opposite. I will then use this online calculator to calculate the two-tail P value for both evaluation metrics to see if each metric is statistically significant by being greater than or less than the α level of 0.05.

Gross Conversion:

Total Days	Total Successes	Two-tail P value	Statistically Significant?
23	19	0.0026	Yes

Net conversion:

Total Days	Total Successes	Two-tail P value	Statistically Significant?
23	13	0.6776	No

To recap, the hypothesis for Gross conversion was that of a decreases in the experiment group compared to the control group. This is due to the hope that the warning about time would deter those who could not commit the minimum of 5 hours. In addition, the hypothesis for Net conversion was to not have a significant decrease in the experiment group compared to the control group (ideally, there would be an increase). As a result, while Gross conversion proved the experiment hypothesis correct, Net conversion did not.

A quick note on the Bonferonni correction. The Bonferonni correction is mainly used to adjusts *p* values due to the increased risk of a Type I Error (false positives) when making multiple statistical tests. In addition, while it it used to reduce the chance of a Type I Error, it increases the chance of a Type II Error (false negatives). For this particular experiment, I need both of the evaluation metrics to pass in order for me to make a recommendation to launch. This increases the chance of a Type II Error, but the Bonferonni correction is meant to reduce the chance of a Type I Error. As a result, I decided not to use the Bonferonni correction, because it is not applicable in this case.

Make a Recommendation

Given the results of my analysis, I would **NOT** recommend a launch at this point. While Net conversion did not end up being practically significant in the negative direction, the negative dmin value was included in the CI. This makes it a bit of a judgment call. While there are signs that this change could have a negative impact on Net conversion (and Udacity’s bottom line), it is not certain. More data is likely needed.

In order to improve this experiment, I would first dig deeper into the data to see I there are any patterns that will shed light on why Net conversion decreased as much as did. It could also lead to ideas I could use in a follow up experiment. It might be necessary to rerun the experiment as is to get more power, or to modify and rerun the experiment. An addition to this experiment could be to add (in addition to the time requirement) a reminder of the prerequisites required for the course. This would force the person to weigh both the necessary time commitment and skill-set needed to be successful in the course.

Follow-Up Experiment: How to Reduce Early Cancellations

An idea that came to mind has to do with the user experience after they have started the free trial. The experiment would start, of course, with splitting the users into control and experiment groups. The users in the experiment group would have discrete pop-up messages show periodically reminding the student of how to get help if they are stuck. The message could say something like this:

> “Having trouble? Don’t hesitate to visit the forums and get help from our mentors and coaches!”

Hypothesis: Giving the students reminders that they are not alone and have somewhere to turn to when they get stuck will encourage them to find the help they need so they do not drop out from frustration.

Unit of Diversion: User-id

- Chosen because it is the easiest (and most invariant) unit that can be used to differentiate between users once they have made accounts.

Metrics:

- **Number of user-ids:** number of users enrolled in the free trial.
- **Number of forum page-views:** number of forum page-views by unique user-ids.
- **Page-view-probability:** number of forum page-views divided by the number of user-ids.
- **Retention:** 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.

Invariant Metrics:

- **Number of user-ids**
  - Chosen as the means with which to keep track of how many users are in each group. This should not be significantly different between the control and experiment groups.

Variant Metrics:

- **Page-view-probability**
  - Chosen to keep track of which group is more likely than the other to visit the forums.
- **Retention**
  - Chosen to see which group is more likely to continue with the course after the free trial ends.

Citations

- https://discussions.udacity.com/t/final-project-stuck-in-calculating-standard-deviation/183317
- https://discussions.udacity.com/t/final-lesson-duration-and-exposure-no-idea-how-to-progress/163802/2
- https://discussions.udacity.com/t/should-retention-be-evaluation-metric/200951/10
- https://discussions.udacity.com/t/effect-size-tests-caluclation-of-gross-conversions-confidence-interval/27181
- https://discussions.udacity.com/t/struggling-with-final-project-effect-size-tests/34035
- https://discussions.udacity.com/t/practical-significance-boundary/218562
- https://discussions.udacity.com/t/bonferroni-correction/201344/25
- http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full
- http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/type-i-and-type-ii-error/
- http://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf
- http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf
- https://discussions.udacity.com/t/could-not-understand-reviewer-comments/227397