# Experiment Overview: Free Trial Screener

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.

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.

## Unit of diversion

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.

# 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.

- **Number of cookies:** That is, number of unique cookies to view the course overview page. ($d_{min}$=3000)
- **Number of user-ids:** That is, number of users who enroll in the free trial. ($d_{min}$=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 triggered). ($d_{min}$=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. ($d_{min}$=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. ($d_{min}$= 0.01)
- **Retention:** That is, 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 user-ids to complete checkout. ($d_{min}$=0.01)
- **Net conversion:** That is, 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. ($d_{min}$= 0.0075)

**Invariant Metrics:**

1. Number of cookies
   Number of cookies will be a good population sizing invariant because it is randomly assigned between control and experiment group.

2. Number of clicks
   Since this is captured even before the user experiences the change, we do not expect these metrics to move. This can be a good invariant metric.

3. Click-through-probability
   Since this is captured even before the user experiences the change, we do not expect these metrics to move. This can be a good invariant metric.

**Evaluation Metrics:**

1. Gross conversion
   Suitable for evaluation metric since the numerator is captured after the user experiences the change. We expect these metrics to move. We expect to see a drop in this metric.

2. Retention
   Suitable for evaluation metric since the numerator is captured after the user experiences the change. We expect these metrics to move. We expect to see an increase in this metric.

3. Net conversion
   Suitable for evaluation metric since the numerator is captured after the user experiences the change. We expect these metrics to move. We expect to see a drop in this metric.

In case of number of user-ids, we cannot use this as an invariant metric since we expect this to change. We cannot use this as an evaluation metric either since this is an absolute number.

**Results required in order to launch the change:**
The goal of the change is to (1) reduce the number of enrollments by unprepared students (or frustrated students), (2) without significantly reducing the enrollments by those students who pass the free trial and complete the course.

- Goal 1 can be measured by Gross conversion metric. Since users will be apprised of the course workload, students who wouldn't be able to meet minimum expectations will not enrol, reducing the number of checkouts. We wish to see a decrease in this metric.

- Goal 2 can be measured by Retention or Net conversion metrics. We expect that the change should not affect the group of students who would remain enrolled after the free trial and complete the course. We wish to see these metrics increase or at least remain unchanged..

## Measuring Variability

Below is the table containing rough estimates of the baseline values for the metrics.

| Metric | Baseline Value |
|---|---|
| Unique cookies to view course overview page per day | 40,000 |
| Unique cookies to click "Start free trial" per day | 3,200 |
| Enrollments per day | 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 |

For binomial distribution with probability $p$ and sample size $N$, the analytical estimate of variability (standard deviation) is given by,

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

Given a sample size of 5,000 cookies visiting the course overview page, below are the analytical estimates of variability for the evaluation metrics.

| Evaluation Metric | Analytical Estimate of Variability |
|---|---|
| Gross conversion | 0.0202 |
| Retention | 0.0549 |
| Net conversion | 0.0156 |

We know that, analytical estimate of variability is close to the empirical estimate when the unit of diversion and unit of analysis are the same. For our experiment, the unit of diversion is cookie, hence the analytical and empirical estimates of variability for Gross conversion and Net conversion metrics would have similar values.

But, for Retention metric, since the unit of analysis (user-ids) is different from the unit of diversion (cookie), the analytical estimate of variability will be much higher. In this case, it is better to calculate the empirical estimate of variability as well.

## Sizing

### Number of Samples vs. Power

While designing an experiment, determining the sample size is a fundamental step. We need to size the experiment sufficiently for a given metric, in order to achieve desired significance level and statistical power.

For our set of metrics, we are not applying Bonferroni correction since these are highly correlated metrics and Bonferroni would be too conservative and would inflate the sample size.

For this experiment, we are using this sample-size calculator to determine the sample size for each metric.

| Metric | Baseline Conversion | Min. Effect Size | Level of Significance | Power | Estimated Sample Size |
|---|---|---|---|---|---|
| Gross conversion | 0.20625 | 0.01 | 0.05 | 0.8 | 25,835 |
| Retention | 0.53 | 0.01 | 0.05 | 0.8 | 39,115 |
| Net conversion | 0.1093125 | 0.01 | 0.05 | 0.8 | 27,413 |

Now that we know the sample size required for each metric in order to reduce the number of type I and type II errors while concluding the results, let's calculate the number of pageviews required for each metric in order to collect the required sample size.

**Baseline sample data:**

Unique cookies to view course overview page per day: 40,000
Unique cookies to click "Start free trial" per day: 3,200
Enrollments per day: 660

**Calculating pageviews per click and enrollment:**

Overview pageviews for each click: 12.5
Overview pageviews for each enrollment: 60.61

Below table shows the number of overview pageviews required for each metric to reach the corresponding estimated sample size.

| Metric | Estimated Sample Size | Pageviews per Unit | No. of Samples Required | Total Pageviews Required |
|---|---|---|---|---|
| Gross conversion (probability of enrolling, given click) | 28,835 | 12.5 | 2 | 6,45,875 |

| | | | | |
|---|---|---|---|---|
| Retention<br>(probability of payment, given enroll) | 39,115 | 60.61 | 2 | 47,41,212 |
| Net conversion<br>(probability of payment, given click) | 27,413 | 12.5 | 2 | 6,83,525 |

Note that we are collecting two samples here, one for the control group and the other for the experiment group.

### Duration vs. Exposure

Now that we know the number of pageviews to the site (overview page) per day and required number of pageviews, let's determine how many days we need to run the experiment in order to obtain statistically significant results.

| Metric | Pageviews per day | Total Pageviews required | Exposure Level | Duration to Collect (days) |
|---|---|---|---|---|
| Gross conversion | 40,000 | 6,45,875 | 100% | 16.15 |
| Retention | 40,000 | 47,41,212 | 100% | 118.53 |
| Net conversion | 40,000 | 6,83,525 | 100% | 17.13 |

Already at 100% exposure level, we can see that we need to run the experiment for 119 days in order to conclude the results for the Retention metric. It is not feasible to run an experiment for such a long time period. But, if we drop this metric, we would only need to run the experiment for 18 days at an exposure level of 100% in order to statistically conclude the results. It is practical to drop this metric rather than running the experiment for a longer time period.

So, after dropping the Retention metric, if we divert only 50% of the site traffic for our experiment, we need to run the experiment for 35 days to collect the required data.

There is no or minimal risk of conducting the experiment at Udacity. There is no risk of users getting hurt or any other long term negative effects.

# Experiment Analysis

## Sanity Checks

Let's do sanity checking for the chosen invariant metrics in order to understand if there are any differences between control and experiment groups.

Given below are the results of sanity checking for each invariant metric at 95% level of confidence.

| Metric | Lower Bound | Upper Bound | Observed Value | Result (Pass/Fail) |
|---|---|---|---|---|
| No. of cookies | 0.4988 | 0.5012 | 0.5006 | Pass |
| No. of clicks on "Start free trial" | 0.4959 | 0.5049 | 0.5005 | Pass |
| CTP on "Start free trial" | -0.0013 | 0.0013 | -0.0001 | Pass |

The observed values for all the chosen invariant metrics are within the 95% confidence interval values, hence we can say that there are no significant differences between control and treatment groups. We can safely go ahead and analyze the results of our experiment.

## Result Analysis

### Effect Size Tests

Let us construct the confidence intervals at 95% confidence level for each of our validation metrics and compare them with statistical level of significance and practical level of significance values in order to draw conclusions. This is a parametric way of testing the hypothesis.

| Metric | Lower Bound | Upper Bound | Min. Effect Size | Statistically Significant? | Practically Significant? |
|---|---|---|---|---|---|
| Gross conversion | -0.0291 | -0.0120 | 0.01 | Yes | Yes |
| Net conversion | -0.0116 | 0.0019 | 0.0075 | No | No |

Statistical significance is determined by whether or not the constructed confidence interval contains 0.

Practical significance is determined by whether or not the constructed confidence interval contains the established minimum effect size ( $d_{min}$ ).

We can see that the difference in Gross conversion metric is both statistically significant and practically significant. But, the difference in Net conversion metric is neither statistically significant nor practically significant.

### Sign Tests

Let's do a sign test in addition to the parametric method so as to double check the experiment results. Sign test is a non-parametric way of testing the hypothesis. We can use this calculator to obtain the below results.

| Metric | No. of Success Outcomes | No. of Trials | p-value from Sign test | Statistically Significant? |
|---|---|---|---|---|
| Gross conversion | 19 | 23 | 0.0026 | Yes |
| Net conversion | 13 | 23 | 0.6776 | No |

Again, we have results similar to the effect size hypothesis tests.
The difference in Gross conversion metric is statistically significant, but the difference in Net conversion metric is not statistically significant.

### Summary

We have seen that both the metrics (Gross conversion and Net conversion) were giving consistent results with Effect size test and Sign test. At 5% significance level and 80% power,

**Gross conversion:**

- Effect size test
  - Statistically significant? - Yes
  - Practically significant? Yes
- Sign test
  - Statistically significant? - Yes

**Net conversion:**

- Effect size test
  - Statistically significant? - No
  - Practically significant? No
- Sign test
  - Statistically significant? - No

## Recommendation

We had two goals to be met in order to confidently launch the change, (1) reduce the number of enrollments by unprepared students (or frustrated students), (2) without significantly reducing the enrollments by those students who pass the free trial and complete the course.

We saw that there was a statistically and practically significant decrease in Gross conversion metric. The “Free trial screener” indeed reduced the number of enrollments by students after displaying the course expectations before checkout. This is what was expected, hence, goal (1) was met.

However, the Net conversion was neither statistically nor practically significant. It isn't behaving like we expected in order to launch the change. Even more, the lower bound of the confidence interval is below the negative boundary of the practical significance. It's possible that this number went down and reduced the number of students who complete the free trial and make at least one payment. In fact, this is not desirable. Hence, goal (2) was not met.

In conclusion, I would not recommend launching the change at this stage. We might have to slice and dice the data in order to clearly understand the impact of the change on users and investigate the reasons. This would help us steer the direction for the follow up experiment.

## References


1. Project instructions
   https://docs.google.com/document/u/1/d/1aCquhIqsUApgsxQ8-SQBAigFDcfWVVohLEXcV6jWbdI/pub?embedded=True

2. Baseline values
   https://docs.google.com/spreadsheets/d/1MYNUtC47Pg8hdoCjOXaHqF-thheGpUshrFA21BAJnNc/edit#gid=0

3. Experiment data
   https://docs.google.com/spreadsheets/d/1Mu5u9GrybDdska-ljPXyBjTpdZIUev_6i7t4LRDfXM8/edit#gid=0

4. Analysis and calculations
   https://github.com/xalemsharan/online_A-B_test/blob/master/ProjectResults.xlsx

5. Online calculator for sizing experiment
   https://www.evanmiller.org/ab-testing/sample-size.html

6. Online Sign and binomial test tool
   https://www.graphpad.com/quickcalcs/binomial1/