# Udacity's A/B Test for Free Trial Screener

## Experiment Overview

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, it was 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.

## Key hypothesis

- There will be a decrease in the number of students who complete the checkout and continue to free trial.
- There will be no significant decrease in the number of students who continue past free trial and complete the course.

## Experiment Design

### Metric Choice

#### Invariant Metrics

Selected invariant metrics: number of cookies, click-through-probability.

Before running the experiment it is necessary to choose invariant metric to perform consistent check during the experiment. A number of cookies will be a good first invariant metric, as soon as the calculations will be performed on the course overview page before implemented change.

For here the number of cookies instead of user-id were selected because users are able to watch videos and take quizzes without creating an account on a site.

In addition, click-through-probability was selected as another invariant metric to evaluate the number of users who generally clicked the “Start free trial” button and make sure is wasn’t affected by new implementation. This also will allow seeing changes in users behavior if it was affected  by any secondary factor.

For here the number of clicks could be the good invariant metric also. But we already have click-through-probability which is basically a normalized number of clicks. So these two metrics are highly related. So, in this case, that would be enough to select only one metric to measure clicks.

The number of user-ids also won’t be a good invariant metric, because we are tracking users by their id only after completing the checkout and enrolling, so we are expecting that the number of stayed for free trial users would be different. This metric can’t be invariant.

### Evaluation Metrics

Selected evaluation metrics: gross conversion, net conversion.

According to initial hypothesis, the newly implemented feature will reduce the number of users that left free trial without continuing education, so the first evaluation metric is Gross conversion: the number of user-ids to complete checkout boundary divided by the number of unique cookies to click the “Start free trial”. The newly implemented feature will filter some amount of users that tend to drop learning after the free trial. We are expecting to see the decrease in the number of users who actually pass the checkout.

We also expect that this change won’t significantly reduce the number of students who pass the free trial and eventually complete the course. In other words, 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 (or Net conversion) should remain without significant decrease. To continue with the free trial the user’s information is asked and the user is tracked further with user-id.

The Retention also could be a good evaluation metric, as soon as it measures 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. But running this metric could be not practically significant as we will have to run the experiment for a really long time, as this metric is associated with the number of user-ids, but not with pageviews. Calculation the duration for this metric showed that the experiment must continue at least 119 days, which is the really long period of time (not really a few weeks). According to this the previous two metrics will be a good evaluation for this experiment.

Another possible evaluation metric to measure is number of user-ids. But this one is not very good as soon as the number of users may be different between control and experiment groups and that can skew the results. Also this metric is just a count, we already picked gross conversion which is ratio and can be more reliable as it is a normalized number of users, but not a raw number.

During the experiment, we expect the following changes in selected evaluation metrics: reducing of gross conversion and having no significant reduce in net conversion. In the case of both metrics meet the expectations there will be possible to make a conclusion of a success of the experiment.

## Measuring Standard Deviation

The decision of using empirical or analytic variability for selected metrics is based on the unit of analysis and the unit of diversion. For both my selected invariant metrics the unit of diversion and unit of analysis is the same: unique cookie. So the analytical estimate of standard deviation tends to be near the empirical estimate of standard deviation.

To estimate the standard deviation of each evaluation metric the baseline values are used:

| **Description** | **Value** |
|---|---|
| Unique cookies to view page per day: | 40000 |
| Unique cookies to click "Start free trial" per day: | 3200 |
| Enrollments per day: | 660 |
| Probability of enrolling, given click (gross conversion): | 0.20625 |
| Probability of enrolling, given click (net conversion): | 0.1093125 |

The sample size is 5000 cookies visiting the course overview page. The appropriate number of units were recalculated according sample size. We are considering that all events are independent and there are only two types event result: click or no click. So for this case the distribution will be binominal. So the STD can be approximated by SE.

As it was discussed earlier for here the analytic estimate of standard deviation is calculated according to the following values:

| **Description** | **Value** |
|---|---|

| | |
|---|---|
| Unique cookies to view page per day: | 5000 |
| Unique cookies to click "Start free trial" per day: | 400 |
| Enrollments per day: | 82.5 |

| **Metric** | **Probability** | **STD** |
|---|---|---|
| Gross conversion: | 0.20625 | 0.0202 |
| Net conversion: | 0.1093 | 0.0156 |

Talking about probability we can use normal distribution (with increasing number of units in sample). For both cases we computing the probability, so for these cases we compute the analytic variability. According to we use normal distribution it is expected that empirical estimate will be comparable from analytic estimate.

## Sizing

### Number of Samples vs. Power

For this case I won't use the Bonferroni correction as soon as I expect it would be too conservative, because my metrics can be correlated. In case the change will have the effect that will cause that one of selected metric to move and probably all metrics will move together.

To compute the number of pages that is necessary to run the experiment properly we need to compute the number of pages for each metric and select the largest number. Also we will need the same amount of pages for both: experiment and control group.

According to $\alpha = 0.05$ and $\beta = 0.2$ the following results were calculated:

| | **Gross conversion** | **Net conversion** |
|---|---|---|
| Sample size per variation | 25,835 | 27,413 |
| Pageviews per group | 322937.5 | 342662.5 |
| Days to run per group | 16.5 | 17.5 |

In this case we need to have 685325 page views for experiment and control groups.

### Duration vs. Exposure

According to previous calculations (number of pageviews), expected unique cookies per day (40000) and diverting the 50% of traffic to the experiment the duration is estimated by 35 days.

In case we divert the 100% of traffic for this experiment it will take only 18 days, which is less that month. The experiment itself doesn't look very risky as soon as we're not dealing with any sensitive data, in this case, the whole traffic can be sent to experiment. Reducing the percentage of traffic will increase the duration, but it will allow us to reduce the number of users that will be affected in case there are any bugs in implemented change. Speaking only about the time it is more preferable to have an 18 days duration, we should also consider if this experiment is delaying others, the shorter period will help us to receive results as soon as possible, but it also makes it risky if there were any errors in new implemented feature.

# Experiment Analysis

## Sanity Checks

Before calculation the actual results of the experiment it is necessary to perform the sanity checks to make sure that invariant metrics didn't experience significant changes.

| | **Experiment group** | **Control group** |
|---|---|---|
| Pageviews | 344660 | 345543 |
| Clicks | 28325 | 28378 |

**Number of cookies:**
Observed p = cont_pageviews / (cont_pageviews + exp_pageviews) = 0.5006
SE = sqrt(0.5 * (1 - 0.5) / (cont_pageviews + exp_pageviews)) = 0.0006018
Margin of Error (α = 0.05) = 1.96 * SE = 0.0011796
CI = (0.4988, 0.5012)

**Click-through-probability:**
Pooled p = (cont_clicks + exp_clicks) / (cont_pageviews + exp_pageviews) = 0.082154
Observed p = exp_clicks / exp_pageviews + cont_clicks / cont_pageviews = 0.00005 ≈ 0
SE pooled = sqrt(p pool * (1 - p pool) / (cont_pageviews + exp_pageviews)) = 0.00066
Margin of Error (α = 0.05) = 1.96 * SE = 0.0000566
CI = (-0.0013, 0.0013)

Both invariant metrics passed sanity checks. Since the fraction in the control group in the confidence interval there is no reason that something went wrong. As soon as both metrics pass the sanity check there is no reason to dig into day-by-day data to figure out if something wrong.

## Result Analysis

### Effect Size Tests

For now there are calculations to figure out does the experiment had effect and are these results are statistically and practically significant.

| | Experiment group | Control group |
|---|---|---|
| Clicks | 17260 | 17293 |
| Enrollments | 3423 | 3785 |
| Payments | 1945 | 2033 |
| Gross conversion | 0.1983198146 | 0.2188746892 |
| Net conversion | 0.1126882966 | 0.1175620193 |

**Gross conversion:**
Pooled p = (cont_enrolls + exp_enrolls) / (cont_clicks + exp_clicks) = 0.2086
SE pooled = sqrt(p pool * (1 - p pool) / (cont_clicks + exp_clicks)) = 0.00437
Margin of Error (α = 0.05) = 1.96 * SE = 0.00857
D-hat (difference) = -0.02055
CI = (-0.0291, -0.012)
This metric is statistically significant as the confidence interval does not include 0. It is also practically significant as soon d min = 0.01 and the confidence interval is much lower so we can see the large enough decrease.

**Net conversion:**
Pooled p = (cont_pays + exp_pays) / (cont_clicks + exp_clicks) = 0.115
SE pooled = sqrt(p pool * (1 - p pool) / (cont_clicks + exp_clicks)) = 0.0034
Margin of Error (α = 0.05) = 1.96 * SE = 0.0067
D-hat (difference) = -0.00487
CI = (-0.0116, 0.0019)
For this metric there is no statistically significant result as the confidence interval does include 0. And it's also not practically significant because the boundary d min = 0.0075 is out the confidence interval.

### Sign Tests

Using day-by-day data I've performed a sign test:

| | Gross conversion | Net conversion |
|---|---|---|
| Days of experiment | 23 | 23 |
| Number of "success" (exp > cont) | 4 | 10 |
| P value | 0.0026 | 0.6776 |

For the gross conversion metric the result is statistically significant with α = 0.05 and boundary = 0.01.

### Summary

Regarding experiment result the gross conversion metric showed significant decrease in number of students that past the checkout and enrolled free trial period. But the net conversion relatively initial hypothesis (there will be no decrease) showed not expected results, so the implemented checkout significantly reduced the number of students who enrolled and past the free trial period. This result doesn't follow the initial hypothesis.
In order to launch the experiment, we need both metrics meet our expectations. But as a result we see that the net conversion does not. In case then both metrics should match the expectations to meet experiment requirements, we can't use the Bonferroni correction here. The Bonferroni correction is used for type I error when it is necessary at least one metric shows expected results. As soon as we expect gross conversion and net conversion follow initial expectations this is type II error, when both metrics must be relevant. According to results of the experiment the gross conversion showed to be the statistical and practical significant. However the net conversion became neither statistically, nor practically significant.

## Recommendation

The following recommendations are built on two metrics: gross conversion and net conversion. After experiment it was discovered that the gross conversion is practically and statistically significant. So as the result we will have less people who will continue with the free trial. But for net conversion the results ended up with neither statistically nor practically significant. As it was discussed earlier there is a decrease in number of students who pass through free trial and continue after payment. Also the confidence interval for net conversion includes the negative of the practical significance boundary, so this decrease in students can be risky for business if launch the change as is. That is not acceptable risk for launch, so it is necessary to run further experiments.

# Follow-Up Experiment

As soon as the main purpose of the Udacity is not only providing the materials for the courses but also improving students experience. I think the main reason for quitting the program is when students expectations don't follow the reality during the education. Starting the studying for each course there are time estimations for each section and project according to specific time per week that student must spend in "classroom". In case the student spend less time for studying and break all deadlines that may cause the early cancellations before completing the course.

Tracking the time student spend in classroom per week may clear all moments when all deadlines were failed. Also if use the results of the previous experiment there is a statement to spend specific time for learning during the week. If not - follow the estimated time for each course. If student spend less time the popup message can be shown in a classroom in the second part of a week. This can be some kind of friendly reminder to focus on education and follow time limits for each section. In case the student spend committed time or more the popup should be not shown.

The main assumption here will be that showing friendly reminder will increase the time students spend in classroom and thus reduce the number of early cancellations, which is the main purpose of the experiment. The unit of diversion in this case will be the user-ids because we are going to run this experiment after user enrolled. Also, the number of user-ids must be considered as invariant metric, as all students start with enrollment. The evaluation metric here will be retention, the probability of payment given enroll. The null hypothesis is there will be no increase in number of payments after free trial period. To consider that experiment was successful it is necessary to reject the null and achieve the statistical and practical significant positive change in retention. In this case we consider that student's expectations are follow the reality and new implementation will reduce the number of early cancellations.

## Resources


- https://www.udacity.com/course/ab-testing--ud257
- https://en.wikipedia.org/wiki/Variance
- https://en.wikipedia.org/wiki/Binomial_distribution
- https://en.wikipedia.org/wiki/Standard_error
- http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf
- http://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf
- http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full