A/B Testing for Udacity's Free Trial Screener

Thuy Quach

1. 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, 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.

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:

Which of the following metrics would you choose to measure for this experiment and why? For each metric you choose, indicate whether you would use it as an invariant metric or an evaluation metric. 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 trigger). (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, 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)
- 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)

You should also decide now what results you will be looking for in order to launch the experiment. Would a change in any one of your evaluation metrics be sufficient? Would you want to see multiple metrics all move or not move at the same time in order to launch? This decision will inform your choices while designing the experiment.

In order to launch the experiment, I want to see that students who complete the checkout after seeing free-trial screener will be less than who do not see it. In other words, Gross Conversion will be statistically decreased. It would be nice to see the Net Conversion will be statistically increased though it is not required. There are number of reasons to remain enrolled after 14-day trial such as quality of class, underestimation of how much time to study new class, etc.

2. Experiment Design:

2.1 Metric Choice:
List which metrics you will use as invariant metrics and evaluation metrics here.

Invariant metrics: Numbers of cookies, Numbers of clicks, Click-through-probability
Evaluation metrics: Gross conversion, Retention, Net conversion.

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

First let's take a look at invariant metrics and evaluation metrics definitions:

- *Invariant metrics* are the metrics that should not change across the experiment and control. Invariant metrics happens before the user see the experiment therefore independent from it. For example: the same of users across countries, languages; the same distribution.
- *Evaluation metrics* are usually the business metrics, like for example market share, number of users, or user experience metrics. Evaluation metrics are dependent on the experiment.

From the definition I could see that

- Number of cookies: is number of unique cookies to view the course overview page. It is a good invariant methods because it is should be the same across the experiment and control. Also, the course overview page is viewed before the experiment therefore it is independent with the experiment.
- Number of user-ids: That is, number of users who enroll in the free trial. It is not a good invariant methods because depending on experiment some users will enroll or not. It is not a good evaluation metric because the number of user-ids can be different between experiment and control groups. Also, from business point of view, number user-ids alone can not tell how the screen effect enrollment.
- 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 a good invariant method because it happens before the experiment therefore it is independent with the experiment.
- 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. It is a

good invariant metric because both number of cookies and number of clicks are invariant metrics.

- 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. It is a good evaluation metric because the data is only for experiment. Also, it is a good business metric to evaluate how the screener effects enrollment.
- 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. It seems like a good evaluation metric because the data is only for experiment. However, given the calculation of pageview of 4741212, it would take Udacity too long to run the experiment given its traffic. Therefore, I did not use it as an evaluation metric.
- 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. It is a good evaluation metric because the data is only for experiment. Also, it is a good business metric to see whether the screen increase or reduce Udacity's revenue.

2.2 Measuring Standard Deviation:

List the standard deviation of each of your evaluation metrics.

Gross conversion : 0.0202
Net conversion : 0.0156

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

Gross conversion and net conversion have the unit of analysis as number of cookies which is also the unit of diversion in the experiment. Therefore, gross conversion and net conversion would be comparable to the empirical standard deviations.

2.3 Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately.

I did not use Bonferroni correction

Number of samples: 685275

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment.

Fraction of traffic: 0.8
Length of experiment: 22 days

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

This experiment is not very risky for Udacity and students.
Udacity:

- It does not affect current students.
- It does not cost much to collect and analyze data.
- The experiment is also simple that less likely to have big bugs.

Students:

- The risk is less than minimal risk which defined as the probability and magnitude of harm that a participant would encounter in normal daily life. Entering how many hours of learning will not hurt students.
- How many hours of learning is not a very sensitive information.

3. Experiment Analysis:

3.1 Sanity Checks:

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check.

	Lower bound	Upper bound	Expected	Sanity check
Number of cookies	0.4988	0.5012	0.5006	Pass
Number of clicks	0.4959	0.5041	0.5005	Pass
Click-through-probability on “Start free trial”	-0.0013	0.0013	0.0001	Pass

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data.

All of the sanity checks were passed.

3.2 Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant.

	Lower bound	Upper bound	Statistical significance	Practical significance
Gross conversion	-0.0291	-0.0120	Yes	Yes
Net conversion	-0.0116	0.0019	No	No

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant.

	P value	Statistical significance
Gross conversion	0.0026	Yes
Net conversion	0.6776	No

3.3 Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

The Bonferroni correction is an adjustment made to P values when several dependent or independent statistical tests are being performed simultaneously on a single data set. If I only

want one of my expectations to be met, then I will have Type I error (the null hypothesis is true and you reject it). In this Type I error case, Bonferroni correction is useful because we want to reduce this error.

In this experiment, I have two expectations which are Gross Conversion to be decrease and Net Conversion not to be reduced. In this case I want both (***all***) of my expectations to be meet. This kind of experiment is more likely to have Type II error (the null hypothesis is false and you false to reject it). Therefore, I don't use Bonferroni correction because it will be too conservative.

3.4 Recommendation
Make a recommendation and briefly describe your reasoning.

From the effect test size results, I see that Gross Conversion is negative and is both statistically and practically significance. It is good because that I expected the screener help screening busy students who unlikely to complete the class.

Net Conversion is range from -0.0116 to 0.0019 and either statistically or practically significance. Net Conversion has a negative number which means the experiment would *reduce* the number of students to continue after the trial and finish the course. It does not meet second part of Udacity's hypothesis that the experiment will not significantly reduce the number of students to continue past the free trial and eventually complete the course.

From the overall analysis, I suggest Udacity not to launch the experiment.

4. Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

Knowing how much time to work on a Nanodegree is a good start for student. However, it is also very important to keep students engaged and motivated. Online learning is a lonely and sometime frustrated experience. Knowing there is a assigned coach/academic advisor to help could increase students engagement and motivate them to finish a class.

One good follow up experiment would be assigning a coach to "get to know you" and periodically check in with the students about challenges and progress (optional, if short of time

for the experiment). It would add some cost to Udacity such as paying for the coach but it would significantly increase the student remain enrolled after 14-day trial (Retention).

Null Hypothesis: assigning a coach to welcome and "get to know" student would not significantly increase the student remain enrolled after 14-day trial (Retention)

Unit of diversion: user-ids since after the student enrolls in the free trial, they are tracked by user-id from that point forward.

Metrics choice:
Invariant metrics: number of user-ids
Evaluation metrics: retention

If our evaluation metric (Retention) is positive and both statistically and practically significance, we can launch the experiment.