P7-Design an A/B Test

Experiment Overview: Free Trial Screener

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.

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.

Experiment Design

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

There are two kind of metrics that measure how the experiment group is better than control. These are invariant metrics and evaluation metrics.

Choice of invariant metrics

Invariant metrics are those which should not change across the experimental and control groups,so it can provide a way to double check the sanity of the experiment design after we conduct the experiment.Because the screener is triggered after clicking “Start free trial” button,the number of clicks,click through probability and pageviews remain the same.The number of user ids cannot be used as invariant metric since it comes after screener is

triggered.So under this setup,we choose the following metrics knowing that the intervention is occuring after the metrics are measured.Therefore,it is clear that the experiment will not have any direct effect on the following metrics:
1.**Number of cookies:**Number of unique cookies to view the course overview page.This is also a population sizing metric and should be equally divided between the experimental and control groups
2.**Number of clicks:** The number of unique users/students (unique cookies) that click on the "start free trial" button.This happens before free trial scanner is triggered
3.**Click-through-probability:** The number of unique cookies to click on the "start free trial" button divided by the number of unique cookies to view the course overview page.

Choice of evaluation metrics
Evaluation metrics are expected to change over the experiment with differences observed between the experimental and control groups.Each of the evaluation metrics are associated with minimum difference(dmin),which is observed among other values to take the final decision regarding whether or not to launch the change.So,in this setup,anything after the screener change shows up will be different across the experimental and control groups.Here are the choices for evaluation metrics:
1.**Gross conversion:** The number of user-ids that enroll in the free trial divided by the number of unique cookies to click on the "start free trial" button.We expect that this metric is lower for experimental group since on clicking the “Start Free Trial” button,the screener triggers up and it helps to filter out those students who are not ready to make time commitment.But in the control group,there is no such filtering process
2.**Retention:** The number of user-ids that stayed enrolled past the 14 day free trial (made a payment) divided by the number of user ids to complete checkout..For this metric too,for the control group,this rate might be lower since they enroll without looking at the time commitments and hence unable to continue the courses.So this metric could actually measure if the screener change had an effect on the retention rate.
3.**Net conversion:**The number of user-ids to remain enrolled past the 14 day free trial (made a payment) divided by the number of unique cookies that clicked on the "start free trial" button.

Unused metric:user-id
The user ids cannot be used as an invariant metric since it is tracked only after student enrols in the free trial.So ,it will be different across the control and experiment groups.It is also not a good evaluation metric since there is no denominator and hence,cannot be normalized.

If our hypothesis is correct,we can see the changes in our evaluation metrics.We expect the gross conversion to be lower since the number of students will be filtered by the screen trigger.And also,regarding the retention rate,our requirements state that it should not the number of students who continue past the 14 day trial and make payment,so retention should not change.Finally,the net conversion should not decrease since number of students who continue past free trial and complete course are unaffected.

Retention is not a part of the launch criteria here because it would require many days to record this metric.It involves the students who remain enrolled and make a payment after 14 day boundary.

Measuring Variability

Baseline table

Unique cookies to view 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

For Bernoulli distribution with probability p and population N, the analytical standard deviation is computed as `std = sqrt(p * (1-p) / N)`.

For 5000 pageviews,

Number of clicks=5000 x 0.08=400

Number of enrollments=5000 x 0.08 x .20625 =82.5

By using the formula of standard deviation,we get

Analytical standard deviation for gross conversion= $\sqrt{\frac{.20625(1-.20625)}{400}}$ =.0202

Analytical standard deviation for retention= $\sqrt{\frac{0.53(1-0.53)}{82.5}}$ =.0549

Analytical standard deviation for net conversion= $\sqrt{\frac{0.1093125(1-0.1093125)}{400}}$ =.0156

Evaluation metric	Standard deviation

Gross conversion	.0202
Retention	.0549
Net conversion	.0156

For gross conversion and net conversion,the denominator is number of cookies and since the unit of diversion is also number of cookies,we can say that analytical variance is comparable to the empirically estimated variance.But,for retention,the denominator is number of enrollments or user-id.So,empirical variance would be higher than analytical variance.

Sizing

Number of Samples vs. Power

The Bonferroni correction was not used in this case.

Given α=0.05 and β=0.2

- Gross conversion(Base conversion rate=20.625%,dmin=1%)
- Retention(Base conversion rate=53%,dmin=1%)
- Net conversion(Base conversion rate=10.93125%,dmin=.75%)

The sample size to power the experiment can be calculated using Evan Miller

- Gross conversion:25835 clicks for each group
- Retention:39115 clicks for each group
- Net conversion:27413 clicks for each group

Click/pageview ratio=3200/40000=0.08

Enrollment/pageview ratio=660/40000=0.0165

Multiplying the sample sizes by unit/page view ratio,we get the total number of pageviews.

Evaluation metric	Minimum detectable effect	Sample size	Unit/Pageview ratio	Number of page views
Gross conversion	.0202	25835	.08	645875
Retention	.0549	39115	.0165	4741212
Net conversion	.0156	27413	.08	685325

Therefore,the number of page views(largest sample size) to conduct the experiment is 4,741,212.

Duration vs. Exposure

We need to specify an exposure based upon the risk involved in the experiment.Here,only minimal risk is involved since no sensitive data is collected about the participants nor there is any physical harm as consequence.Therefore,100% exposure is safe and can be used.

If we use Gross conversion,Retention and Net conversion as evaluation metrics and the daily pageview baseline value is 40,000 then

Number of days=4741212 /40000=119

This duration is unreasonably long for an A/B testing experiment.So,we eliminate retention as an evaluation metric.

Next,using only gross conversion and net conversion,

Number of days=685325/4000 = 18

The experiment will now require 18 days to complete.

Experiment Analysis

Sanity checks

- Control group:#pageviews=345543   #clicks=28378
- Experiment group: #pageviews=344660   #clicks=28325

Confidence level= 95%

1.**Number of cookies**

Standard deviation= $\sqrt{\frac{0.5\,(0.5)}{345543+344660}}$ =0.0006018

Margin of error,m=1.96 x SD=1.96 x 0.000602 =0.00118

∴Confidence interval=(0.49882,0.50118)

Observed= $\frac{345543}{345543+344660}$ =0.5006

2.**Number of clicks**

SD= $\sqrt{\frac{0.5(0.5)}{28738+28325}}$ = 0.0021

Margin of error,m=1.96 x 0.0021 = .004116

Confidence interval =(0.495884,0.504116)

Observed = $\frac{28378}{28378+28325}$ =.500467

Similarly,the confidence intervals for click through probability(CTP) are also computed.

Summary of the computed confidence interval and whether or not the metric pass the sanity check

Invariant Metric	Lower bound CI	Upper bound CI	Observed	Pass/Fail
Number of Cookies	.49882	.50118	.5006	Pass
Number of clicks	.4959	.5041	.5005	Pass
CTP	.0812	.0830	.0822	Pass

Since click-through-probability is the number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page and both number of clicks on "Start free trial" and number of cookies pass sanity check,click-through-probability also pass the sanity check.

Result Analysis

- Control group:#Clicks = 17293 #Enrolment = 3785 #Payment = 2033
- Experiment group:#Clicks = 17260 #Enrolment = 3423 #Payments = 1945

Effect Size Tests

Gross conversion

Observed value = $\frac{3423+3785}{17293+17260}$ =0.2086

$$SD = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{N_1}+\frac{1}{N_2})}$$

$$= \sqrt{.2086(1-.2086)(\frac{1}{17293}+\frac{1}{17260}}\ )$$

=.004372

m=1.96 * 0.004372=.008569

Observed difference= $\frac{3423}{17260}$ - $\frac{3785}{17293}$ =-0.02055

Net conversion

Observed value= $\frac{1945+2033}{17293+17260}$ =.1151

SD= $\sqrt{(.1151(1-.1151)(\frac{1}{17293}+\frac{1}{17260}}$))=.003434

m=1.96 x .003434 =.00673

Observed difference = $\frac{1945}{17260}$ - $\frac{2033}{17293}$ = -.00487

Evaluation metric	Upper bound CI	Lower bound CI	Statistically or practically significant
Gross conversion	-.012	-.0291	Yes(Statistically and practically)
Net conversion	.0019	-.0116	No(Statistically and practically)

Sign Tests

Using the day-by-day data,the sign test is performed.We evaluate each day of the experiment to see if there is a positive or negative difference across the experimental and control groups.Each positive difference can be counted as success and each negative difference,a failure.

Evaluation metric	No: of success	No; of trials	Probability	Two-tailed p value	Statistically significant
Gross conversion	4	23	.5	.0026	Yes
Net conversion	10	23	.5	.6776	No

Summary

The experiment conducted was the potential Udacity students were diverted by cookie into control and experimental groups.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.However,the control group was not asked to do so.The experiment was conducted to determine if filtering the users based on time commitment would improve overall user experience and improve coaches' capacity to support students who are likely to complete the course.The invariant metric chosen were Number of cookies,Number of clicks on “Start free trial” and click-through-probability for purpose of validation and sanity check whereas gross conversion and net conversion were used as evaluation metrics.The null hypothesis is that there is no significant difference in the evaluation metrics between the control and experiental groups.To launch the experiment,the null hypothesis should be rejected for all the evaluation metrics.Also,the difference between two groups should be larger than minimum detectable effect for all metrics.I

I have not used Bonferonni correction in my analysis.The Bonferonni correction is a method for controlling for type I errors (false positives) when using multiple metrics in which relevance of any of the evaluation metrics matches with the hypothesis. In our case,we run the risk of increasing the type I errors when the number of metrics increases.Our requirement must be that all metrics must be relevant to launch the experiment i.e we cannot base our decisions on one metric alone.

On conducting,the sanity check,the expected equal distribution of cookies into control and experimental groups was observed at 95% CI.The difference in Gross conversion between the two groups was found to be statistically significant at 95% CI ,so we reject the null hypothesis.However,Net conversion was found to be neither practically nor statistically significant at 95% CI.

Recommendation

Observations

1.We can see that gross conversion is practically significant and negative,but Net conversion is not statistically significant
2.The 95% CI of net conversion include the negative number of practical significance boundary which points to the decrease in Udacity's revenue

I would recommend against launching the change because:
1.Experiment was successful in reducing the number of unprepared students who enrolled in the free trial as shown by the gross conversion analysis of screener
2.However,the net conversion showed that the experiment reduced the number of students who continue after the 14 day boundary and hence make a payment.

Follow-Up Experiment

I would recommend doing an experiment in which the students who clicked on the 'Start free trial' button for a course would be redirected to a small quiz session which covers the prerequisites for the course.Those students who pass the quiz,say 80% score,can be then taken to the payment portal.For those who do not pass the quiz,few pointers on where to improve can be given and it can also be said that you can start the course when you are ready with the prerequisites.Another option can also be given for the students where they can start the free trial regardless of whether they pass the quiz or not.I believe that taking the quizzes will help the students to know better their skill levels and make a good decision regarding starting the course i.e not enrolling for the course and then becoming frustrated that they are not able to complete it.

The null hypothesis would be that the incorporation of the quiz will not improve the number of students enrolled beyond the 14 day trial period.

The unit of diversion can be cookies.Even though it is less stable than user id,we cannot use user id since it is assigned only after enrolment.

Invariant metric

Number of clicks: Number of unique cookies to click on “Start free trial “ button.Since this happens before the enrolment,it can be used as invariant metric.

An equal distribution between the control and experimental groups can be expected.

Evaluation metric

Retention: If there is a practical and statistical significant increase in retention,we can say change is successful.