Experiment Overview

The Udacity course home pages have two options: 'Start free trial' and 'Access course materials'.

- Clicking the 'Start free trial' prompt the user to enter their credit card information, subsequently enroll in a free trial for the paid version of the course, after which they are automatically charged.
- Users who click 'Access course materials' will be able to view the videos and take the quizzes for free, but receive no coaching support, verified certificate or project feedback.

Change

For this experiment, Udacity tested a change wherein those users who clicked 'Start free trial' were asked how much time they were willing to devote to the course.

- Users choosing 5 or more hours per week would be taken through the checkout process as usual.
- For users indicating fewer than 5 hours per week, a message would appear indicating the need for a greater time to commitment to enable success 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.

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 students experience and improve coaches' capacity to support students who are likely to complete the course.

The primary aim

- Improve the overall student experience
- Improve coaches' capacity to support students who are likely to complete the course

Set up Hypothesis

Null Hypothesis: This approach might not make a significant change and might not be effective in reducing the early Udacity course cancellation.

Alternative Hypothesis:

- This might reduce the number of frustrated students who left 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

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

Invariant metrics
- Invariant metrics should not change across experiment and control groups
- Can provide a way to double check the integrity of the experiment design after the experiment was conducted.

In this experiment, the screener change shows up after clicking on the 'start free trial' button, thus, the number of pageviews, clicks, and the click-through-probability are expected to remain the same for both control and experiment groups.

Therefore, the invariance metrics chosen for this experiment are:
- **Number of cookies**: Number of unique cookies to view the course overview page
- **Number of clicks** : Number of unique cookies to click the start free trial button

Evaluation Metrics

- Evaluation metrics are dependent on the experiment, and are expected to change over the course of the experiment.
- Each evaluation metric is associated with a minimum difference (dmin) that must be observed for consideration in the decision to launch the experiment.

In this experiment, anything after the screener change shows up, Number of user-id, Gross conversion, Retention, Net conversion could be affected for control and experiment groups.

Therefore, the evaluation metrics chosen for this experiment are:
Gross conversion: Number of user-id to complete checkout and enroll in the free trial / number of unique cookies to click the 'Start free trial' button
- It is directly dependent on the experiment and could measure whether or not the screener had an effect on enrollment.
- We expect that the value will be lower in the Experiment group because those users who are likely to drop during the 14-day trial (are not able to commit more than 5 hours per week) would be filtered by the screener.
- For the Control group, users won't see any pop-up message so they will enroll without any consideration of number of hours they can commit per week.
- Therefore, the gross conversion in the control group is expected to be higher than in the experiment group.

Retention: Number of user-ids to remain enroll for the 14-day trial period and make their first payment / number of user-ids to complete checkout
- It is directly dependent on the experiment and could measure whether or not the screener change had an effect on 14-day retention rate.
- We expect the value will be higher in the Experiment group because majority of the users are those who can commit 5 hours per week and are more likely to remain enrolled after 14-day period and start their first initial payment.
- For the Control group, the users enroll for the course without considering the time availability commitment, which might cause lower retention rate.

Net Conversion: Number of user-ids to remain enrolled for 14-day trial and at least make their first payment / Number of unique cookies to click the 'Start free trial' button.
- It is directly dependent on the experiment and could measure whether or not the screener change had an effect on the first payment completion rate.

- It would be great if this value increases after the experiment. But considering the expected decrease in the total number of users enrolling in the 14-day free trial, we expect this value does not decrease significantly after the experiment.

Unused Metrics

Number of User-ID: Number of users who enroll in the free trial.
- This is not appropriate invariant metric because user-ids are tracked only after student enrolling in the free trial, so the number of use-ids might be different between Control and Experiment groups.

Click-through-probability: Number of unique cookies to click the 'Start free trial' button / number of unique cookies to view the course overview page
- This is a good invariant metric since the clicks are occurred before the users see the experiment, therefore it does not depend on our test .
- However, the number of cookies and number of clicks are already sufficient to use as invariant metric, thus this metric may be redundant for analysis.

Metrics Variability

Measuring Standard Deviation
The number of clicks and enrollments follows a Binomial distribution and by the Central Limit Theorem, the distribution of the three rates(Gross Conversion, Retention and Net Conversion) is Gaussian.

Given the daily sample of 5000 cookies, the number of clicks and enrollments can be calculated using the baseline values:
Number of clicks = (5000 x 3200) / 40000 = 400
Number of enrollments = (5000 x 660) / 40000 = 82.5

Table 1. Analytical Estimate of Standard Deviation

Evaluation Metrics	Baseline Value	Standard Deviation
Gross Conversion	0.206250	0.0202
Retention	0.53	0.0549
Net Conversion	0.109313	0.0156

Both Gross Conversion and Net Conversion using Number of cookies as denominator, which is also unit of diversion. Here, the Unit of Diversion and Unit of Analysis are the same, which indicate the Analytical estimate would be comparable to the Empirical standard deviation.

For Retention, the denominator is 'Number of users enrolled the courseware', which is not similar as unit of Diversion. Thus, the empirical variability may be different from analytical estimate, thus we perform both analytical and empirical estimate for the metric

Sizing

Number of Sample vs.Power
Given the type 1 error rate of Alpha equals 0.05, type 2 error Beta equals 0.2, and the Minimum Detectable Effect for each evaluation metric, the Sample Size required to power the experiment

appropriately can be calculated using Evan Miller. Then, the total number of pageviews can be calculated using the given pageview ratio.

Click / Pageview ratio = 3200 / 40000 = 0.08 clicks/pageview
Enrollment / Pageview ratio = 660 /40000 = 0.0165 enrollment/pageview

Table 2. Results of Sample size calculation

Evaluation Metric	Baseline Value	Minimum Detectable Effect	dmin	Sample Size	Number of groups	Total sample size	Unit/ pageview ratio	Total number of pageviews
Gross Conversion	0.206250	0.0202	0.01	25835	2	51670	0.08	645875
Retention	0.53	0.0549	0.01	39115	2	78230	0.0165	4741213
Net Conversion	0.109313	0.0156	0.0075	27413	2	54826	0.08	685325

Based on the results in table 2, a total of 4741213 pageviews is required to conducted the experiment.

Duration vs. Exposure

With daily page view baseline value of 40000, the number of page view for retention would need about 119 (4741213 / 40000) days, even if we divert 100% traffic. It is unreasonably long for an A/B testing experiment.

Therefore, I eliminate Retention as the evaluation metrics. The total number of required pageviews is decreased to 685,325.

Considering that this is not a risky experiment as the change is small and it won't cause too many trouble in the overall business, I choose to direct 70% of the traffic (40000* 0.7 = 28000) to the experiment. Thus, it would takes approximately 25 days (685,325/28000 = 25) to run the experiment.

Experiment Analysis

Sanity Checks

- Having conducted the experiment, each of the invariant metrics need double-check whether the underlying assumptions are being met.
- Cookies and clicks are expected to divided evenly between Control and Experiment groups.
- Using an expected rate of diversion of 0.5, the standard deviation can be calculated and a 95% confidence interval can be constructed around the expected value

Table3. Results of sanity checks

Invariant Metric	Expected value	Observed value	CI Lower	CI Upper	Pass Sanity
Number of cookies	0.5	0.499360	0.497641	0.502359	TRUE
Number of clicks	0.5	0.499533	0.491769	0.508231	TRUE

According to the results of table 3, both invariant metrics, cookies and clicks, pass the sanity check since their observed values are within 95% confidence interval

Result Analysis

Effect Size Tests
For each evaluation metric, statistical and practical significance (whether or not the size of the effect is relevant from a business standpoint) should be tested. The Minimum Detectable Effect is the smallest difference that we will accept between Experiment and Control groups in order to be practically significant.

Using the data collected, we conclude the rate in experiment and Control groups for each evaluation metric (Gross conversion, Net conversion) and then define a new variable, that is, the Difference between the rates (experiment - control). Using this newly define variable, we construct a confidence interval which will then set a range for the expected difference.

Table 4. Result of effect size tests

Evaluation Metric	D-min	Observed diff	CI Lower	CI Upper	Results
Gross Conversion	0.01	-0.0206	-0.0291	-0.0120	Statistically and Practically significant
Net Conversion	0.0075	-0.0049	-0.0116	0.0019	Neither statistically nor practically significant

Since 95% confidence interval does not include zero and the minimum detectable effect value, Gross Conversion is both statistically and practically significant. In terms of Net conversion, the 95% confidence interval includes zero and the minimum detectable effect value, indicating neither statistically nor practically significant.

Sign Tests
A binomial sign test will be conducted to further test each of the evaluation metrics. Each day of the experiment is evaluated to see if there is a positive or negative difference across groups (experiment - control). Each positive difference is counted as a success, and each negative difference as a failure. Comparing the resulting p-values for each metric to determine significance.

Table 5. Results of sign test

Evaluation Metric	# of success	# of trials	Probability	Two-tail p-value	Results
Gross Conversion	4	23	0.5	0.0026	Statistically significant
Net Conversion	10	23	0.5	0.6776	Not Statistically significant

According to the result of table 5, Gross conversion rate has 4 of 23 success for a two-tailed p-value of 0.0026 indicating statistical significance of Gross Conversion. Net Conversion has 10 of 23 success and a two-tailed p-value of 0.6776 indicating that Net Conversion is not statistically significant. Both are consistent with the hypothesis test results.

Summary

- In this experiment, potential Udacity users were diverted by cookies into either control group or experiment group. After clicking 'Start free trial' button on the home page, users in the experiment group were asked how much time they are willing to devote to the course, while users in the control group were not.
- This experiment was designed to determine whether filtering users as a function of studying time commitment would improve the overall user experience and improve coaches' capability to support users who are likely to complete the course, without significantly reducing the number of students who continue past the free trial.
- Number of cookies, Number of clicks on 'Start free trial' button were chosen as invariant metrics while Gross Conversion (enrollment/ cookie), Net Conversion (payment/ cookie) were chosen as evaluation metrics.
- The null hypothesis is that there is no significant difference in the two evaluation metrics between control and experiment groups. In order to launch the experiment, the null hypothesis must reject for all evaluation metrics, as well as the differences between two groups should larger than the minimum detectable effect for each evaluation metric.
- Because the evaluation metrics in the experiment have high correlation and all evaluation metrics must have statistically significant differences, thus the Bonferroni correction will be too conservative and will not be used during the analysis phrase.

- The sanity test results revealed, for the two invariant metrics, the expected equal distribution of cookies into control and experiment group at the 95% confidence interval.
- In the term of the effect size hypothesis test, the difference in Gross Conversion between the control and experiment group was both statistically and practically significant and the null hypothesis was rejected. However, the difference in Net Conversion was neither statistically nor practically significant at 95% confidence interval

Recommendation

Based on the analysis results, I do not recommend to launch this experiment. The reason are as follows:

1. Although Gross Conversion turned out to be negative and practically significant, Net Conversion results are both statistically and practically insignificant, which did not meet the acceptance criteria that null hypothesis must be rejected for all evaluation metrics
2. The 95%CI of Net Conversion does include the negative number of the practical significant boundary, which suggests the risk of hurting Udacity's business - decrease revenue.