Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here.

Experiment: When “start free trial” is clicked, students will be prompted and ask how much time they can commit to the course. If student selects 5 or more hours per week, then he/she will be taken to the normal checkout process. However, if he/she selects less than 5 hours, then a warning would be prompted asking student to spend at least 5 hours per week on the course. Student has an option to continue enrolling to the “free trial” or access the free materials instead.

Initial hypothesis:

a. “Experiment warning prompt 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”

b. ” without significantly reducing the number of students to continue past the free trial and eventually complete the course. “

The following invariant metrics are used: **Number of cookies**, **Number of clicks**, **Click-through-probability**

The following evaluation metrics are used**: 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.

Number of cookies: The number of cookie measures the number of unique users visiting the course page. Since this should be evenly distributed between control and experimental groups, and user visits the course page before the experiment, this should be used as an invariant metric. This was not used as an evaluation metric since this occurs before the experiment and is independent of it

Number of clicks: The number of clicks measures the number of cookies that click the start free trial button. Since this occurs before the experiment also, this should be used as an invariant metric. This was not used as an evaluation metric with the same reason of number of cookies. This occurs before the experiment and is independent of it

Click-through-probability: The clicks occurred prior to users seeing the warning prompt. Therefore, this is not dependent on our test and it should be used as an invariant metric. Because of its independence of the experiment, this was not used as an evaluation metric.

Gross conversion: Gross conversion measures the number of users who enrolled in the free trial over number of users who clicked the “start free trail” button.. This should be used an an evaluation metric since it can be used to check if the experiment makes a difference in the enrollment.

The assumption is that for the controlled group, no warning will be prompted, where in the experiment group, user would be prompted a warning about time commitment. Since the controlled group has no prompt, it is expected to have higher gross conversion. This can be used as an evaluation metric to check if the experiment makes a significant difference in the enrollment.

Empirical analysis vs analytical analysis – denominator = # of cookies. Since # of cookies is also the unit of diversion, and equivalent to the unit of analysis, therefore, the analytical estimate should be comparable to the empirical variability

Retention: Measure the number of users remained enrolled for trial period and made payment over number of users who enrolled in free trial. In the experiment group, users are aware of the time commitment, since they’re being prompted with the warning. Retention rate for the experiment group expects to be higher. For the control group, since they were not given the warning prompt, cancellation rate could be higher compared to the experimental group. Therefore, we’re using this as an evaluation metric

The assumption for this is that retention rate in the experimental group would be higher, since users would be prompted with warning on hours commitment. Therefore, they’re more prepared to spend at least the minimum required hours on the course, leading to higher retention rate. On the other side, users in the controlled group do not get warning. Perhaps some of them do not spend minimum number of hours on the course and might lead to lower retention rate in the controlled group. We’ll use this to confirm part A of our hypothesis.

Empirical analysis vs analytical analysis – denominator is number of users enrolled in the course, which is different from # of cookies (unit of diversion). Since the unit of analysis is not the same as the unit of diversion, the analytical and empirical estimates are different

Net conversion: Net conversion measures the number of users remained enrolled for trial and also made payment over the number of users clicked the start free trial button.

The assumption is that in the experiment group, students are aware of the time commitment after the warning. They’re more likely to remain and pay. Where in the controlled group, they’re not aware of the time commitment and are more likely to opt out of the trail program. Therefore, we can use this as an evaluation metric and use this to confirm part B of the hypothesis. While it would be nice to see an increase in the result for net conversion, an increase is **not required** to launch the experiment. We expect there’s **no decrease** in net conversion

Empirical analysis vs analytical analysis - denominator = # of cookies. Since # of cookies is also the unit of diversion, and equivalent to the unit of analysis, therefore, the analytical estimate should be comparable to the empirical variability
User-id: Measures the number of users enrolled in the trial. It was not used as an invariant metric since it's dependent on the experiment, and we expect to see different value in control and experiment group. So it was not used as an invariant metric. It was not used as an evaluation metric either since the number of enrolled users for different groups could be different and we could not ensure that both groups get about the same number of enrolled users. So this was not used as an evaluation metric either.

Measuring Standard Deviation

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.

Evaluation Metric	Standard Deviation
Gross Conversion	0.0202
Retention	0.0549
Net Conversion	0.0156

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. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

The approach I choose is **NOT to** deploy Bonferroni correction.

	Gross Conversion	Retention	Net Conversion
Baseline Conversion	20.625%	53%	10.93%
Min Detectable Effect	1%	1%	1%
Alpha	0.05	0.05	0.05
Beta	0.2	0.2	0.2
1 – beta	0.8	0.8	0.8
Sample size	25835	39155	27413
Number of groups	2	2	2
Total sample size	51670	78230	54826
cliicks/pageview	3200/40000 = 0.08	660/40000=0.0165	3200/40000=0.08
pageviews	645875	4741212	685325

Since we're taking the largest number for pageviews, 4741212 will be the number of pageviews we need

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. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

Initially, I chose to include retention, which requires 4741212 pageviews. Even if I use 100% of the traffic, it would still take close to 119 days to complete the experiment, given 40,000 pages per day. By removing retention, the experiment will take 18 days instead (100% traffic, 685325 total required pageviews)

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

In lesson 2, "Minimal risk is defined as the probability and magnitude of harm that a participant would encounter in normal daily life. The harm considered encompasses physical, psychological and emotional, social, and economic concerns".

In our experiment, it is very unlikely that a student would be exposed to any harm or gets hurt due the the proposed change. Also, the website is not financial or health related.

Also, we're not dealing with sensitive data such as financial, health related, personal info such as SSN, credit card number, etc. Therefore, this experiment is not considered to be risky and 100% of the traffic can be diverted.

Experiment Analysis

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. (These should be the answers from the "Sanity Checks" quiz.)

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

	Lower bound	Upper bound	Observed	Passes
# of cookies	0.4988	0.5012	0.5006	Yes
Click-through-probability on "start free trial"	0.0812	0.0830	0.0822	Yes

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. (These should be the answers from the "Effect Size Tests" quiz.)

	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. (These should be the answers from the "Sign Tests" quiz.)

	p-value	Statistical significance (alpha = 0.05)
Gross conversion	0.0026	Yes
Net Conversion	0.6776	No

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 experiment was conducted and the students (unit of diversion = cookie) were divided into 2 groups. “Number of cookies”, “number of clicks on start free trial” and “click through probability” were used as invariants metrics. “Gross conversion”, “Net conversion” were selected to be evaluation metrics. Initially, “retention” was also selected as evaluation metric. However, the duration of using that as evaluation metric was too long. Therefore, it was removed so that the duration of the experiment was more reasonable.

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. Therefore, it is more likely that one of the multiple metrics will be falsely positive as the number of metrics increases. We should use Bonferroni correction when the criteria of launching the tests is to go when one of the evaluation metric shows a statistical significance. When the criteria of launching the experiment is that **ALL** metrics need to show statistical significance, then Bonferroni correction is not needed.

Since the acceptance criteria requires both statistically significant differences for all evaluation metrics, Bonferroni correction should not be used.

Recommendation

Before the experiment started, the goal was to have a decrease in gross conversion and no decrease in net conversion.

The result for gross conversion was negative and practically significant. This is expected. However, the result for net conversion showed that the confidence interval includes the negative of the practical significance boundary. Since the second part of the hypothesis is "without significantly reducing the number of students to continue past the free trial and eventually complete the course", therefore, this is not an acceptable risk for the business to take and the recommendation is not to proceed with the change.

Follow-Up Experiment

Further step can be to design a short course which can be completed within the 14-day trial period, requiring students to spend 10 hrs/week.

The benefits that the student gain from this short course include real course taking experience and if he/she can commit 10 hrs per week for 2 weeks. If student cannot commit 10 hrs per week, this experience can give student a much more realistic time that he can commit to taking the course.

Setup	When clicked on "start free trial", experiment group is asked to complete a short course by the end of the free trail
Null hypothesis	Students successfully completing short course by end of free trial are more likely to remain in the program and make their first payment
Unit of diversion	User-id – since this occurs after user clicks on "start free trial", click-based and cookie-based metrics are not good. User-id should be used instead to keep track
invariant metrics	User-id
Evaluation metrics	**Retention Rate**: retention rate measures the number of users remained enrolled for trial period and made payment over number of users who enrolled in free trial. We choose this as evaluation metric since it's not cookie/click based and is user-id based. If a statistically and practically significant positive change in retention is the result of the follow-up experiment, we would then recommend