Experiment Design

Metric Choice

Invariant metrics:

Number of cookies, Number of clicks, Click-through-probability.

Evaluation metrics:

Gross conversion, Net Conversion.

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. The metrics happened before students are asked should be the equal for control and experiment group. Therefore, "**Number of cookies", "Number of clicks", "Click-through-probability"** are proper invariant metrics.

Number of user-ids, that is number of users enroll in the free trial. Because the experiment aims at reducing the number of students who left the free trail for lack of time, the metric is not invariant. The metric is neither a good evaluation metric. The metric would vary with the number of clicks per day. Fractional metrics would be better choice.

Retention, that is the probability of payment, given enroll. As expected from the experiment, if the students without enough time choose to access the course instead of enrolling, the enrollment rate would fall. And the number of students who continue past the free trial and make payment doesn't change, so the retention should increase. It looks like a good evaluation metrics. But the metric will need a large number of pageviews, which needs about 119 days of experiment. That's pretty long time running, so retention is not selected as evaluation metrics.

Gross conversion and **net conversion** are good evaluation metrics. Gross conversion is the probability of enrolling given click, and net conversion is the probability of payment given click. The number of clicks is invariant, but the probability of enrolling and probability of payment would change due to the experiment. To launch the experiment, we would expect the experimental group shows lower gross conversion than the control group, above the minimum detectable effect 1%. At the same time, the net conversion should remain the same for experimental group and control group, within minimum detectable effect 0.75%.

Measuring Standard Deviation

Analytic estimate of standard deviation, given a sample size of 5000 cookies visiting the course overview page:

Metrics	Standard deviation
Gross conversion	0.0202
Net Conversion	0.0156

The analytic estimate would be comparable to the empirical variability, because the unit of division and unit of analysis are both 'cookie'.

Sizing

Number of Samples vs. Power

Bonferroni correction will not be used.

For metric Gross conversion, baseline conversion rate=0.20625, minimum detectable effect 1%, alpha=0.05, beta=0.2, it will need a sample size of 25835/0.08*2 = 645875.

For metric Gross conversion, baseline conversion rate=0.1093125, minimum detectable effect 0.75%, alpha=0.05, beta=0.2, it will need a sample size of 27413/0.08*2 = 685325.

So 685325 pageviews will be needed to power the experiment.

Duration vs. Exposure

I would like to divert 100% Udacity traffic (50% to control group, 50% to experiment group) to this experiment, and it will need 18 days to run the experiment. By allocating all the traffic, it will need relative short time for the experiment. And Udacity can make decisions earlier whether to launch the change and have a better allocation of coaches resources. The experiment is generally of low risk, it just add a screen asking the committed hours of students. But there is the risk that the change may keep out the students who may find the course interesting during the free trial and decide to spend more hours. 18 days is a proper length for conducting the experiment.

Experiment Analysis

Sanity Checks

For each of your invariant metrics, 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	Observed	Sanity
Number of cookies	0.4988	0.5012	0.5006	Pass
Number of clicks	0.4959	0.5042	0.5005	Pass
Click-through-probability	0.0812	0.0830	0.0822	Pass

Result Analysis

Effect Size Tests

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

	Lower bound	Upper bound	Stat. significant	Practical significant
Gross conversion	-0.0291	-0.0120	Yes	Yes
Net conversion	-0.0116	0.0019	No	No

Sign Tests

	“Success”/Trials	p-value	Stat. significance
Gross conversion	19/23	0.0026	Yes
Net Conversion	13/23	0.6776	No

Summary

Bonferroni correction is not used. Because we use two metrics to evaluate the experiment. And we need both of the metric to meet our criteria. Bonferroni is not applicable in the "ALL" case.
In our case, it will reduce number of valid launches.
No discrepancy is found between the effect size hypothesis and the sign test.

Recommendation

I would recommend not to launch the experiment.
The gross conversion for the experiment group is lower with both statistically and practically significant. That's as expected that in experiment group, some of students without enough time are guided not to enroll the free trial. But for the net conversion, the lower bound -0.0116 is lower than –dmin of -0.0075. It's risky that launching the experiment may decrease the net conversion. Further test of the net conversion will be needed.

Follow-Up Experiment

To reduce number of frustrated students who cancel early in the course, I would like to design a following experiment:
After users are enrolled in the free trail, they are provided with an interesting and fancy pre-start project. The project doesn't require much pre-knowledge, but requires at least 10 hours highly devoted time. If the users could finish the project, then they are provided with the 50% tuition back coupon (the one currently used at Udacity) for the following contents in the course. The hypothesis is that users will build confidence and find pleasure in the prestart project, and then feel less frustrated. They may be more willing to pay to continue, because the provided coupon is they earned with efforts. Therefore, the number of frustrated students who cancel early may decrease.
The unit of diversion is user-id. Because the students are tracked by user-id after they enroll in the course.
The invariant metric is number of user-ids, that is the number of users who enroll in the free-trial. No changes are made before users' enrollment, so the metric should be same for control group and experiment group.
The evaluation metric is number of students who pay divided by number of user-ids. The hypothesis is that the experiment can encourage students to continue and pay the course, which leads to a higher value of the metric.