Experiment Design

Metric Choice

Invariant Metrics:

Number of cookies:
The free trial page is changing, this should not affect the number of users landing on this page(as they haven't seen it) and hence this metric should not have any effect on the test and is hence an invariant metric.

Number of clicks
This metric does not depend on how the start free trial page is rendered.

Click through probability
The users have not seen the start free trial page before they decide the click on the button, hence this is an invariant metric.

Evaluation Metrics:

Gross Conversion:
This is the 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.
This is an evaluation metric as the 'Start free trial' page with the **time estimate** influences whether a student will sign up or not.

Net Conversion:
This is the 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.
This is also an evaluation metric as these numbers can very well be affected by the time estimates given in in the free trial page.

Retention:
This is the 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.
This is also an evaluation metric as the new free trial page with the time estimates can very well have an impact on the number of users who decide to stay past the 14 day free trial and pay for the course.

Neither:

Number of user ids:
This is the number of users who enroll in the free trial. We should be checking the ratio of the number of users who enroll in the free trial and the number of users who click on the start free trial button which is the Gross Conversion metric. Therefore this metric doesn't make sense to be used for our a/b test.

Launch Criterion:

I will be looking at the results of my evaluation metrics for my control and experiment groups.

The goal is to reduce enrollment by unprepared students without harming revenue. The first part is measured by gross conversion. The second part is measured by net conversion. This could go up and that would be great but seeing no decrease should be sufficient (so can go up or stay the same).

Gross Conversion: If the experiment goes well, we should see an increase in the number of students that complete checkout and enroll in the free trial in our experiment group(as they will see the pop up) as compared to our control group. This is the most important metric to check in our test as this metric isolates the pop up feature's effectiveness.

Retention: If the experiment goes well, we should see an increase in the number of user-ids to remain enrolled past the 14-day free day trial in the experiment group.

Net Conversion: This could go up or stay the same as our objective to ensure that revenue is not harmed but that could mean lower sign ups at the start but higher longevity in terms of the duration of time the student sticks with the program.

Measuring Standard Deviation

Evaluation Metric	**Standard Deviation**
Gross Conversion	.0202
Retention	.0549
Net Conversion	.0156

Gross Conversion:

Both the unit being used for analysis and the unit of diversion is the cookie, and hence the analytic variance should be close to the empirical variance.

Net Conversion:
Similar to Gross Conversion, the unit being used for analysis and the unit of diversion is the cookie, and hence the analytic variance should be close to the empirical variance.

Retention:
We could do an empirical test here as the metric is the number of user ids. But considering the amount of time it would take to run the a/b test with Retention as a metric(explained later in the report) I dropped this as an evaluation metric.

Sizing

Number of Samples vs. Power

Evaluation metric	Baseline conversion rate	d_min	Sample size needed	Number of page views needed
Gross Conversion	20.625%	.01	25,835	645,875
Retention	53%	.01	39,115	4,741,212
Net Conversion	10.93125%	.0075	27,413	685,325

I used a tool to calculate the number of samples needed. I did not use Bonferroni correction as the metrics being used in this test are correlated and because both of the evaluation metrics are required to satisfy launch criteria thereby reducing the probability of type 1 error. If either of the evaluation metrics were enough to satisfy the launch the criteria, then the probability of a false positive or type 1 error goes up and hence would make sense to apply the correction which is not the case here.
Our null hypothesis rejection requires that both of our tests show statistical significance in order to reject our experiment null hypothesis.
We reject our null hypothesis if gross conversion were to go down in a statistically significant way, while net conversion were to remain unchanged. We would not reject on a single metric in this experiment, hence we do not use the Bonferroni correction.

Based on the data I obtained on the number of page views needed, it would take over 118 days to run the test if i included Retention as an evaluation metric and considering that this is too long a duration, I am dropping it as an evaluation metric.

Duration vs. Exposure

Number of days needed for each evaluation metric is as follows:

Gross Conversion -> 17 days, Net Conversion -> 18 days and Retention -> 118 days.

I will be directing 100% of the traffic towards this a/b test with a 50% split each for the control and experiment groups. If the rest does not run very well it will affect only half the user base and if that is the case we can stop the test. If we reduced our audience, we would have to run the test for a longer period, which is not ideal for our a/b test. Therefore the total number of page views needed is 685,325. This is a low risk experiment as we are not dealing with sensitive data of the user like medical/financial/private data that could cause harm. Hence, we are safe with directing 100% of the traffic towards this experiment.

Experiment Analysis

Sanity Checks(classroom link)

Metric	Lower Bound	Upper Bound	Observed	Passes
Number of Cookies	.4988	.5012	.5006	yes
Number of clicks on 'start free trial'	.4959	.5041	.5005	yes
Click through probability	.0812	.0830	.0821	yes

Result Analysis

Effect Size Tests(classroom link, link 2)

Metric	Lower Bound	Upper Bound	Statistical Significance	Practical Significance
Gross	-.0291	-.0120	yes	yes

Conversion				
Net Conversion	-.0116	.0019	no	no

As Net Conversion metric included zero, it is not statistically significant and not practically significant either.

Sign Tests

Metric	p value	Statistical Significance
Gross Conversion	.0026	yes
Net Conversion	.6776	no

Summary

I did not use the Bonferroni correction as our null hypothesis rejection requires that both of our tests show statistical significance in order to reject our experiment null hypothesis.
We reject our null hypothesis if gross conversion were to go down in a statistically significant way, while net conversion were to remain unchanged. We would not reject on a single metric in this experiment, hence we do not use the Bonferroni correction. If either of the evaluation metrics were enough to satisfy or dissatisfy the launch of the feature, then the probability of a false positive or type 1 error goes up and hence would make sense to apply the correction which is not the case here.
I did not see any discrepancies between the effect size hypothesis and the sign test.

Recommendation

As the gross conversion went down, it shows that there are relatively fewer people enrolling(completing checkout after clicking on 'start free trial' which translates to a lower cost for udacity as it has to support fewer students but at the same time we don't want our net conversion rate(students staying past 14 days after clicking on 'start free trial' to be affected by much which it wasn't (p score of .6776 (higher the p score, the more likely we are to reject our null hypothesis that the new feature does not affect enrollment)). The confidence interval includes the negative of our practical significance boundary for net conversion is -0.0075(-0.75% change) which means it could have a practical significance on the business.
Therefore, considering our goal here is to increase the number of paying students at Udacity while reducing churn, I'd saw we have to run the experiment again with more power.

Follow-Up Experiments

A follow up experiment could be a discount that is applied when students click on the 'start free trial' button to students based on the completion of the first project within a deadline.

The hypothesis would be that more students are encouraged to enroll in the program if they see a discount applied to their account when they sign up.

The unit of diversion would be the user_ids as they are the students that sign up.

The evaluation metric would be the 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.

The invariant metric would be the number of user_ids as the total number of use ids wouldn't change unless we added an additional SEO initiative.

Another experiment could be a 'rolling' discount that would essentially be a discount that applied to the student after completing each project. So instead of the 50% back on tuition at the end of the program, a student would get a discount on their fee that would increase, as the student completed more projects. For example: If a student completes P1, a 20% discount is applied thereafter, after completing P2, it goes to 30% and so on.

The unit of diversion would be the user_ids as they are the students that sign up.

The evaluation metric would be 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.

Similar to the previous experiment, the invariant metric would be the number of user_ids.