A/B Test Design

Metric Choice

Choosing Invariant Metrics:

Chosen:
Number of cookies - There is no difference between the control and experiment when trying to view the course overview page, therefore the number of cookies should be the same in each group.

Number of clicks - There is no difference between the control and experiment before they click the 'Start free trial' button, therefore the number of clicks should be the same in each group.

Click-through-probability - Since there is no difference in the 'number of cookies' and 'number of clicks' metrics between the experiment and the control, there will be no difference in this metric.

Not Chosen:
Number of user-ids - The experiment group is expected to have fewer users enroll in the free-trial, therefore consistency is not expected between the groups.

Gross conversion - The experiment group is expected to have fewer user-ids enroll in the free trial, therefore consistency is not expected between the groups.

Retention - The experiment group is expected to have more user-ids remain past the free-trial period and have fewer enroll in the free-trial, therefore consistency is not expected between the groups.

Net conversion - The experiment group is expected to have relatively more user-ids remain enrolled past the free-trial period, therefore consistency is not expected between the groups.

Choosing Evaluation Metrics:

Chosen:
Gross conversion - Fewer user-ids are expected to complete the checkout and enroll in the free trial in the experiment, but the number of unique cookies to click the "Start free trial" button should be the same. Therefore, the gross conversion ratio should be higher for the control than the experiment. To launch the experiment, we need the experiment value to be at least 0.01 lower than the control value.

Retention - At least as many user-ids in the experiment are expected to remain enrolled past the free-trial period, and fewer are expected to complete the free-trial checkout. Therefore, the retention ratio is expected to be higher for the experiment than the control. To launch the experiment, we need the experiment value to be no more than 0.01 less than the control value.

Net conversion - At least as many user-ids in the experiment are expected to remain enrolled past the free-trial period, and the number of unique cookies to click the “Start free trial” button should be the same. Therefore, the net conversion ratio should not decrease for the experiment. To launch the experiment, we need the experiment value to be no more than 0.0075 lower than the control value.

Not Chosen:
Number of cookies - There is no difference between the control and experiment when trying to view the course overview page, therefore the number of cookies should be the same in each group.

Number of user-ids - Although it is a valid evaluation metric, I am not choosing it because it is not normalized and if we have slightly different sized experiment and control groups, its accuracy will decrease. Therefore, it is not as useful for comparison as the chosen evaluation metrics.

Number of clicks - There is no difference between the control and experiment before they click the ‘Start free trial’ button, therefore the number of clicks should be the same in each group.

Click-through-probability - Since there is no difference in the ‘number of cookies’ and ‘number of clicks’ metrics between the experiment and the control, there will be no difference in this metric.

Measuring Standard Deviation

Gross Conversion = 0.0202
Analytical estimate should be comparable to the empirical variability because the unit of analysis and unit of diversion are the same: cookie.

Retention = 0.0549
I do not expect the analytical estimate to be comparable to the empirical variability because the unit of analysis and unit of diversion are different. Unit of analysis: user-id. Unit of diversion: cookie. I expect the analytical estimate to underestimate variability, as a user-id is unique to an individual, but an individual can use many cookies.

Net conversion = 0.0156
Analytical estimate should be comparable to the empirical variability because the unit of analysis and unit of diversion are the same: a cookie.

Sizing

Number of Samples vs. Power

The retention metric requires the greatest number of page-views, 4,741,212. However, even if 100% of traffic was diverted towards the experiment, the experiment would last for 118 days. Given this excessive length, we are forced to no longer use retention as an evaluation metric. The net conversion metric requires the next greatest number of page-views, 685,325. This will be the number of page-views that we will use.

Duration vs. Exposure

I would divert 100% of traffic towards this experiment, which would result in the experiment lasting for 18 days. This is not a risky experiment as it does not expose participants of the experiment to any additional risk. Participants are only answering a simple question about how much time they think they can dedicate to the course, therefore we are not dealing with any sensitive data. In addition, it is important to have low risk experiments completed as quickly as possible so that costs are not wasted by prolonging the experiment, and to have traffic freed up for future experiments.

Experiment Analysis

Sanity Checks

Invariant Metric	Lower Confidence Interval (95%)	Upper Confidence Interval (95%)	Observed Value	Sanity Check
Number of Cookies	0.4988	0.5012	0.5006	Passes
Number of Clicks	0.4959	0.5041	0.5005	Passes
Click-through-probability	0.0812	0.0830	0.0822	Passes

Result Analysis

Effect Size Tests

Evaluation Metric	Lower Confidence Interval (95%)	Upper Confidence Interval (95%)	Statistically Significant	Practically Significant
Gross Conversion	-0.0291	-0.0120	Yes	Yes
Net Conversion	-0.0116	0.0019	No	No

Sign Tests

Evaluation Metric	p-value	Statistically Significant
Gross Conversion	0.0026	Yes
Net Conversion	0.6776	No

Summary

We do not need to use the Bonferroni correction because it would increase the chance of a type II error. To launch our experiment we need all metrics to be satisfied, therefore we want to limit the probability of false negatives.

There were no discrepancies between the effect size tests and the sign tests.

Recommendation

Given the results of the effect size tests and sign tests, I recommend that the experiment not be launched in its current form. Although the gross conversion metric is both statistically and practically significant, the net conversion metric did not satisfy the required criteria.

One thought for why the net conversion metric may have failed, is that students see that they should put at least five hours of work towards the course each week. Seeing this may 'anchor' their expectations (i.e. establish a benchmark) that five hours per week is enough to complete this course in a sufficient amount of time. During the trial, students may work for five hours in a week, realize how little progress they have made, think the course will take far longer to complete, and decide to quit.

To avoid the scenario above from happening, I propose two slight variations to the experiment:

1. After students click 'start free trial', they will be prompted with the question, "Do you think that you will *regularly* be able to commit 10 hours per week to this course?" Students will select yes or no, then advance to the checkout if they selected yes, or be recommended to use the free courses if they selected no.
    - 10 hours per week could become a benchmark for the students, and using the word 'regularly' should help the students to not feel overwhelmed that they must work at least 10 hours per week. Compared to the original experiment, student may work more with this new benchmark, make more progress during the free-trial period, not become as frustrated with their lack of progress, and remain enrolled in the course.
2. When students are prompted with either the 5 hour or 10 hour question, they will be asked a second question, "This course can be challenging and you may feel frustrated at times. Do you agree to work through these difficulties to accomplish your goal of completing the course?" Students must check the "I can do it!" box to proceed to the checkout.

These two alternatives could be tested with the same methodology as the current experiment.

Follow-Up Experiment

Experiment Overview: Motivational Quotes

To keep students motivated and willing to continue their courses, despite being frustrated, those in the experiment group will be prompted with a motivational/inspirational quote after answering every third quiz correctly. Students should be rewarded for their efforts of correctly answering quizzes, and we do not want to overrun the students with quotes, which is why they will only be presented after every third quiz. Students in the control will not be prompted with any quotes, but will continue to receive the regular encouragement prompts, "Well Done!", "Great Job!", etc.

The hypothesis is that students who receive the motivational quotes will be inspired to continue with their course even when they are discouraged and want to quit before finishing. The unit of diversion is the students' user-ids because this will be unique to each student and they must be signed in to complete the course, and therefore receive the motivational quotes.

Choice of Metrics

Enrollments: The number of user-ids that enroll in a course.
Retention: The number of user-ids that remain enrolled past the free-trial, thereby making at least one payment, divided by the number of 'Enrollments.'

Reasons for Metrics

Enrollments: To be used as the invariant metric as there is no difference between the experiment and the control when students enroll in a course.
Retention: To be used as the evaluation metric, which will show an increase in value in the experiment, should the motivational quotes be having an affect.