Experiment Design

Metric Choice

Invariant metrics:

- Number of cookies - The experiment is at a stage after the cookie gets assigned. So this should not be affected by the experiment and should stay invariant.
- Number of clicks - users click on Start Trial before the question in the experiment. Hence must be invariant
- Click-through-probability - As number of cookies and number of clicks both are invariant , this is also independent of the experiment.

Evaluation metrics: All these are dependent on the experiment and hence

- Gross conversion - This is the ratio of number of enrollments to total number of clicks . Ideally we expect this ratio to be less than the pre-experiment value , as the experiment reduces the number of students who may not be able to spend at least 5 hours a week.
- Retention - Ratio of number of students who at least make one payment to those who enrolled - enrollment depends on the experiment and depending on the retention value (if greater than pre-experiment value) , we can comment if the experiment helped to reduce the number of unreliable / possible drop out enrollments
- Net conversion - Ratio of number of students who enrolled and stayed after 14 days to the total number of students that clicked on the “Start Free trial” - This also depends on the experiment due to the number of enrollments. If the values after the experiment are less than the values prior to the experiment , we can say that it helped reduce the number of enrollments that have a high probability of dropping out.

Number of Userids - This cannot be an invariant metric as it depends on the enrollments . And there is no guarantee to have same number of user ids all days and hence can’t be an evaluation metric due to the skew. As there is no denominator , the metric is not normalized.

To launch the test ,
Gross Conversion should decrease indicating that we successfully reduced the number of enrollments eventually reducing our costs - practically significant decrease
And
Net conversion / Retention should increase (or at least remain the same) indicating that our revenue has increased or has not been affected negatively by the experiment - no statistically significant decrease.

Measuring Standard Deviation

Gross conversion: 0.0202
Retention: 0.0549
Net conversion: 0.0156

In the cases of Gross Conversion and Net conversion , the unit of diversion and Unit of Analysis are same and hence their analytical and empirical estimates will be comparable. Retention on the other hand needs the number of users enrolled and may have different analytical and empirical estimates.

Sizing

Number of Samples vs. Power

Although we are using multiple metrics , I have not used Bonferroni correction as these did not seem to be independent metrics and applying Bonferroni may lead to a high rate of false negatives .

From Evan Miller's website , sample sizes as below
For Gross Conversion alpha=20.625% , dmin=1% Samplesize = 25835
For Net conversion alpha=10.931 ,, dmin=0.75% Samplesize=27413
For retention alpha 53% dmin=1% Samplesize=39115

We have to use the maximum number of samples . If we use 39115 from Retention , we would need 4741213 clicks accounting to 119 days if we divert the entire traffic. As this does not look feasible, I will ignore retention and proceed with Gross Conversion and Net conversion.

To get 27413 samples at a Click through probablity of 0.08 ,total page views needed = 685325 (for both experimental and control so 2*27413 /0.08)

Duration vs. Exposure

With daily traffic of 40000, I'd direct 75 % of my traffic (30000) to the experiment, which means it would take us approximately 23 days (685325/30000 = 22.84) .

As the experiment does not affect any existing paid students as well as the ones who are determined , I don’t see any risk in diverting majority of the traffic to the experiment. I have not used 100 percent traffic though to ensure that the experiment runs for at least a week more than the free trial period.

Experiment Analysis

Sanity Checks

1. Number of cookies:
Total Control group's pageviews: 345543
Total Experiment group's pageviews: 344660
Total pageview: 690203
Probability of cookie in control or experiment group: 0.5
SE = sqrt(0.5*(1-0.5)*1/(345543+344660)) = 0.0006018
Margin of error = SE * 1.96 = 0.0011796
Confidence Interval = [0.4988,0.5012]
Observed value = 344660/690203 = 0.5006

2. Number of clicks:
Total Control group's clicks: 28378
Total Experiment group's clicks: 28325
Total pageview: 56703
Probability of cookie in control or experiment group: 0.5
SE = sqrt(0.5*(1-0.5)*1/(28378+28325)) = 0.0021
Margin of error = SE * 1.96 = 0.0041
Confidence Interval = [0.4959,0.5041]
Observed value = 28378/56703 = 0.50046

Results:
Number of cookies: [0.4988,0.5012]; observed value .5006 lies with in the confidence interval and hence it passes sanity check
Number of clicks : [0.4959,0.5041]; observed value .50046 lies with in the confidence interval and hence it passes sanity check

Result Analysis
Effect Size Tests

For Gross Conversion , below are the counts:

	Control Group	Experiment Group
Total Number of enrolments	3785	3423
Total Number of clicks	17293	17260
Probability	0.2189	0.1983

SE = 0.004371675385
m = SE * 1.96 = 0.00856848375
Pooled Probability = 0.2086
D hat = -0.02055Confidence Interval = [-0.0291,-0.0120]

Statistically significant as CI does not have 0 and Practically significant as dmin (0.01) is also not in the CI.

For Net Conversion , below are the counts:

	Control Group	Experiment Group
Total Number of payments	2033	1945
Total Number of clicks	17293	17260
Probability	0.1175	0.1127

SE = 0.003434133513
m = SE * 1.96 = 0.0067
Pooled Probability = 0.2086
D hat = -0.0049
Confidence Interval = [-0.0116,0.0018]

CI has 0 as well as the dmin (+/-0.0075) . Hence this is neither statistically significant nor practically significant

Sign Tests

With the help of http://graphpad.com/quickcalcs/binomial1.cfm

Gross Conversion :

Number of days for which experimental values are more than control values : 4
Number of trials : 23
Probability: 0.5
Two-tailed p-value : 0.0026

p-value = 0.0026 < alpha level = 0.025. Therefore the data is statistically and practically significant.

Net Conversion :

Number of days for which payments in experiment group are more than control group : 10
Number of trials: 23
Probability: 0.5
Two-tailed p-value : 0.6776

p-value = 0.6776 > alpha level = 0.025. Hence the net conversion is both statistically and practically insignificant.

Summary

The effect size and sign tests show same results

I have not used Bonferroni correction as the evaluation metrics being used are not entirely independent and the correction would be too conservative . Also as we are not evaluating using each of the metric independently but as our decision depends on the result of both the metrics together , we should not apply the correction in this case.

Recommendation

I do not recommend the launch. From the Gross conversion metric , we have reduced the number of dropouts which will help for better allocation of udacity resources and hence is a positive outcome. But the Net conversion CI has negative numbers and that may represent decrease in revenue which is not a good sign. The negative effects of the screener seem to be more than the positive effects and hence I do not recommend a launch.

Follow-Up Experiment

Based on my personal experience , if Udacity recommends a few prerequisite courses to be completed during the trial period , the probability of students dropping out in between the course will come down . These prerequisite courses will ensure that students have the necessary basics before they opt and also give them a better understanding of the effort needed to complete the degree. This will motivate and prepare students better for a nanodegree which needs them to have patience as well as put efforts consistently over a period of time.
Udacity can provide a rating of this prerequisite course (course + quizzes) as well as the average hours needed to complete based on a student's expertise (For e.g., 2 hours for an expert , 10 hours for intermediate , 20 for beginner etc .,) Also the course should be relevant to the nanodegree the student is looking for and it should be possible for a beginner to complete the course in 6 - 7 days at 2 hours a day .

Null Hypothesis : Adding prerequisite courses does not increase the retention of students .

Unit of Diversion : User Ids . Users have to register on Udacity and complete the prerequisite courses using those userids.

Invariant Metrics : Total Number of Userids, as the experiment is after enrollment , Userids will not get affected by it .

Evaluation Metric : Retention - If the experiment succeeds , number of students that continue after the trial period should increase . This will increase the revenue for udacity.
(I was thinking of Retention but have written Net Conversion)

References :

1. Udacity course on A/B testing
2. Statistics for Engineers and Scientists - William navidi
3. https://en.wikipedia.org/wiki/Bonferroni_correction
4. http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf
5. http://graphpad.com/quickcalcs/binomial1.cfm
6. Evan Miller's statistical tools