Experiment Design

Metric Choice

Invariant Metrics

Invariant metrics does not changes across experimental or control groups.

The metrics chosen as invariant are:
1.Number of cookies
2.Number of clicks
3.Click-through probability

The number of unique cookies remains stable throughout the experiment as the page-views (traffic) remains stable throughout the experiment as well.

The number of clicks on the "start free trial" button will remain stable because the experiment takes place after clicking the button.

Click-through-probability also remains stable as it involves number of unique cookies clicking the button only and the number of unique cookies are made stable hence, CTP should remain stable as well.

UserID is not taken as either invariant or evaluation metric as the it is the count of users after getting enrolled and will be impacted by the change prior to it. And also it is not a normalized value, hence it will be difficult to compare instead Gross conversion will stand as a better evaluation metric in the subject.

Evaluation Metrics

Evaluation metrics are assumed to change over the course of the experiment as according to the hypothesis, the number of user-ids to enroll will be decreased as those students who are not able to fulfill the commitment required for the course gets eliminated and the students who progresses has a clarity over time commitment resulting in payments/checkouts after 14 days not decreasing from control group.

As of successful experiment,

1.Decrease the enrollment of unprepared students, i.e the gross conversion should decrease

2.and not decreasing the number of students who complete the 14 days free trial and make first payment, i.e the net conversion should either stay the same or increase.

Gross conversion: The number of user-ids that enroll in the free trial divided by the number of unique cookies to click on the "start free trial" button. Numerator will get affected as the denominator(number of unique cookies stays same).The number of enrollments are expected to be go down so as the gross conversion, therefore it will be evaluation metric.

Net conversion: The number of user-ids to remain enrolled past the 14 day free trial (made first payment) divided by the number of unique cookies that clicked on the "start free trial" button. Numerator will get affected as the denominator(number of unique cookies stays same). The number of payments are expected to increase so as the net conversion or at least producing a non- negative impact producing an increase in first payment/continued enrollment after the free trial ;therefore it will be an evaluation metric.

Retention: The number of user-ids that stayed enrolled past the 14 day free trial (made a payment) divided by the number of unique cookies that clicked on the "start free trial" button. It will require 4741212 page views to detect a minimal change for retention and it is not practically possible to stretch the experiment that long.Therefore it will not be used as Evaluation metric though it has potential.

Measuring Standard Deviation

The number of clicks and enrollments depicts binomial distribution.Also the denominator for Gross Conversion and Net conversion are unique cookies(Unit of Diversion) which has a value of 5000 for visiting course overview page for each sample, hence the measure of standard deviation is expected very low, hence the empirical value is expected to be close to analytical value for that reason.

Gross Conversion:

sqrt(0.20625*(1-0.20625) / (0.08*5000)) = 0.020231

Retention:

sqrt(0.53*(1-0.53)/ 5000*(660/40000)) = 0.0549

Net Conversion:

sqrt(0.1093*(1-0.1093) / (0.08*5000)) = 0.0156

Whereas for retention metric,the units of diversion and analysis are different (user-id v/s cookies), it is expected that the analytical standard deviation and the empirical standard deviation will have a difference.

Sizing

Number of Samples vs. Power

The Bonferroni Correction was not used as it is too conservative for this experiment. The metrics used for evaluation are highly correlated so the assumption for independent event is also a vulnerable thougth.

Page-views required for Gross-conversion:

25835/ 0.8 * 2 = 645876

Page-views required for Net-conversion:

27413 / 0.8 * 2 = 685325

Standard alpha value of 0.05 and beta value of 0.2.

Hence, the number of Page-views required are 685325 for completing the experiment.

Duration vs. Exposure

The duration of the experiment is chosen 18 days, as it is very low risk campaign.

100 % traffic will be diverted because no sensitive data is being collected, the quickest possible duration is the smartest choice for a low risk campaign.

The experiment seems risk-free from many different angles as it doesn't affect the student currently enrolled and it is not a change that will alter the decision of already interested student to progress as the commitment ask is just an acknowledgement received by the student for even more clarity over the subject, it is an aid in decision making not decision making change.
Also as from technical side there is just one UI change with pop-up that too will not be

any hard task for the engineers, though i recommend to divert a small portion first to test before diverting 100% traffic.

Experiment Analysis

Sanity Checks

Alpha value = 0.05, hence the Confidence interval is of 95%

Invariant Metric	Lower Bound	Upper Bound	Observed	Pass/Fail
Number of Cookies	0.4988	0.5012	0.5006	Pass
Number of Clicks	0.4959	0.5041	0.5005	Pass
Click-through Probability	0.0812	0.0830	0.0822	Pass

All Sanity Checks passed, hence Analysis can be further processed.Calculations are as Follows:

Number of Cookies:

> Observed: Views-Control / (Views-Control + Views-Experiment)
> Confidence interval: 0.5 +/- sqrt(0.5*0.5/(VC +VE)) = +/- 0.0012

Number of clicks on "Start free trial":

> Observed: Clicks-Control / (Clicks-Control + Clicks-Experiment)
> Confidence interval: 0.5 +/- sqrt(0.5*0.5 /(CC + CE) = +/- 0.0041

Click-through-probability on "Start free trial":

> Observed: Clicks-Control / Views-Control
> Confidence interval: 0.0821 +/- sqrt(0.5*0.5/345543) = +/- 0.00085

Result Analysis

Effect Size Tests

Analysis done on 95% confidence interval for the difference between the experiment and control groups.

Evaluation Metric	Lower Bound	Upper Bound	Significance
Gross Conversion	-0.0291	-0.012	Yes (statistically and Practically)
Net Conversion	-0.0116	.0019	No (statistically and Practically)

Gross Conversion :

p_{pool} = (control-enrollments + experiment-enrollments) / (control-clicks + experiment-clicks) = 0.208607

Standard Deviation = sqrt(0.208607 * (1-0.208607)*(1/clicks-control + 1/clicks-experimental)) = 0.004372

Margin of error = 0.004372*1.96 = 0.0085685

D = p_{exp} - p_{cont} = 0.19832 - 0.218875 = -0.02055

Hence,it is practically significant(CI doesn’t include dmin)
And Statistically significant(CI doesn’t include 0).

Similarly,Net conversion not significant either statistically or practically for the same reason.

Sign Tests

Sign test calculated using day-by-day data by calculating p-value from graphpad.com.

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

Null Hypothesis(H_0): There is no statistically significant difference in both Gross conversion and Net conversion between two groups.
Based on the p-values; For Gross conversion we reject the null hypothesis,
And for Net conversion we fail to reject the null.

Summary

I didn't use the Bonferroni correction. The Bonferroni correction is commonly used to adjust p value when making multiple comparisons, because as the number of tests increases, so does the likelihood of false positives. It is very common to use the correction when there is one universal null hypothesis (Ho) for all tests, and any metric with statistical significance would lead to reject the Ho and trigger the launch.

I choose to assess the statistical significance of individual test respectively and check whether or not both metrics are satisfied to trigger a launch. This approach already makes it less likely to detect false positives and even with a single false positive, it can't govern a decision.In addition, gross conversion and net conversion are highly correlated metrics, which means if further correction is applied, it would end up being too hard to detect the true positives and result in increasing likelihood of false negatives. Lowering the power of a test is not what we want, because a single false negative from any of the two tests would not trigger the launch. It is obviously not favored to have such a tradeoff by increasing false negatives in this experiment.

Also, The sign and effect size test both showed gross conversion to be statistically significant, while net conversion is not.

Recommendation

Based on the analysis Udacity shouldn’t launch the change as it is because while gross conversion was statistically and practically significant, net conversion was not in fact the net conversion got adversely affected with the experiment as looking at the Confidence Interval: (-0.0116, 0.0019), the lower bound value is smaller than the negative practical significance dmin = - 0.0075, which implies the Udacity revenue may get hurt. I would suggest performing a retrospective analysis and calculating average number of hours for students who pass the first project vs who don’t and then check with the quick assessment.

Follow-Up Experiment

During the 14 day trial students are left alone to complete the videos and project work without direct interaction from Udacity.Instead via user-IDs, mapping the progress of the new student by tracking the progress made by each student via watching videos,completing quizzes,etc. If a student's progress is behind the pace of the recommended pace by Udacity, then student should be prompted to devote more time by a notification and also asking student if they require coaching assistant for any problem they facing.Also i want to speak about content student goes through in 14 days should be a milestone to achieve to give student an all round feeling of completing a part.

The hypothesis is that a kind reminder early on in the process will help the student in two ways 1) give students a reference check on their progress and 2) provide students with resources allowing them to pick up the path. In turn, this will either eliminate more students who just can't dedicate the time while also encouraging other students to work harder to make the deadlines and avoiding those early cancellations.

I would use the user-id as the unit of diversion, enrolled students only, with the number of user-ids as an invariant metric and number of payments divided by the number of unique user-id's as an evaluation metric.

The number of students who enroll should not change across experimental or control groups so it's a good candidate for an invariant metric, while the number of payments /

completed free trials divided by the number of user-ids should change over experimental and the control group and can therefore be used as an evaluation metric.

References:

Wikipedia, Google, Graph Pad