Experiment Design

Metric Choice

Number of cookies: invariant metric

Since the Course overview page appears ahead of free trial screener in funnel, thus number of cookies should not vary in experiment and control

Number of user-ids: none

Since some of the users revert to 'access free material' option, post viewing the experiment thus User-Ids who enroll in free trial are dependent on the experiment. Hence, it cannot be an invariance metric.
Also, it makes for a poor evaluation metric as it is redundant compared to the other metrics. The number of user-ids or enrolled users can fluctuate a lot with respect to the number of start free trial clicks on a given day, and thus not a good proxy for this experiment

Number of clicks: invariance metric

This metric does not depend on how the start free trial page is rendered, much like the number of cookies

Click through probability: Invariance metric

Similar to number of cookies and clicks, since the users have not seen the start free trial page before they decide the click on the button, the click through probability also is not dependent on the test being carried out

Gross conversion: evaluation metric

The rendering of the start free trial page influences the number of users signing up for the free trial. That is, is the 5 or more hours per week suggestion likely to affect conversion rates - this is one question we would like to understand through this A/B test. Therefore, this is a good evaluation metric

Retention: evaluation metric

Likewise, it can be presumed that prompting users about the 5 or more hours per week will have an effect on the ratio of users who make payments versus those who finish the free trial, and thus making this metric good for evaluation

Net conversion: evaluation metric

The ratio of users who make payment over those who see the start free trial page is dependent on the rendering of that page and the 5 or more hours per week suggestion. Hence, being a good overall goal of the A/B test and a good evaluation metric

I will look at Gross conversion and Net conversion. The first metric will show us whether we lower our costs by introducing the screener. The second metric will show how the change affects our revenues.

I will leave Retention, despite it being evaluation metric, because it would have taken too long to gather data on this to support or deny the stated hypothesis of this A/B test.

To launch the experiment, I will require Gross conversion to have a practically significant decrease, while net conversion will not to decrease statistically significance, which indicate the screener whether or not affect the revenues

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

To evaluate whether the analytical estimates of standard deviation are accurate and matches the empirical standard deviation, the unit of analysis and unit of diversion are compared for each evaluation metric. A Bernoulli distribution is assumed here with probability p and population N where the standard deviation is given by sqrt(p*(1-p)/N).

Gross conversion

p = 0.20625 (given)
N = 5000 * 0.08 = 400
std dev = sqrt(0.20625 * (1-0.20625) / 400) = 0.0202

Here, analytical estimate is used for this evaluation metric as it is likely to match the empirical variance. The reason being, the unit of analysis and the unit of diversion is the cookie

Net conversion

p = 0.1093125
N = 5000 * 0.08 = 400
std dev = sqrt(0.1093125 * (1-0.1093125) / 400) = 0.0156

Net conversion 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. The analytical estimate is likely accurate as both the unit of analysis and unit of diversion are cookies as is with gross conversion.

Sizing

Number of Samples vs. Power

As the metrics used in this experiment are highly correlated, I decided against using the Bonfessoni correction as it will be too conservative in the figures calculated.

Online calculator was used to generate the number of samples needed

Evaluation metric	Baseline conversion rate	d_min	Sample size needed	Number of pageviews needed
Gross conversion	20.625%	1%	25,835	645,875
Retention	53%	1%	39,115	4,741,212
Net conversion	10.93125%	0.75%	27,413	685,325

Using gross conversion and net conversion as evaluation metrics here as retention will make it a long running experiment (117 days for 4,741,212 pageviews).
Thus the required number of pageviews being 685,325

Duration vs. Exposure

Risk – Since this experiment does not ask any sensitive information or display harmful results thus we can consider it as risk-free test. The financial implications are already being measured as part of metric tracking (Net conversion)

Making this a 100:0 test, we can ensure that the experiment being evaluated within 18 days (685325 / 40000)

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)

Sanity Checks

Number of cookies

control group total = 345543
experiment group total = 344660
standard deviation = sqrt(0.5 * 0.5 / (345543 + 344660)) = 0.0006018
margin of error = 1.96 * 0.0006018 = 0.0011796
lower bound = 0.5 - 0.0011797 = 0.4988
upper bound = 0.5 + 0.0011797 = 0.5012
observed = 345543 / (345543 + 344660) = 0.5006
The observed value is within the bounds, and therefore this invariant metric passes the sanity check.

Number of clicks on "start free trial"

control group total = 28378
experiment group total = 28325
standard deviation = sqrt(0.5 * 0.5 / (28378 + 28325)) = 0.0021
margin of error = 1.96 * 0.0021 = 0.0041
lower bound = 0.5 - 0.0041 = 0.4959
upper bound = 0.5 + 0.0041 = 0.5041
observed = 28378 / (28378 + 28325) = 0.5005
The observed value is within the bounds, and therefore this invariant metric passes the sanity check.

Click-through-probability on "start free trial"

control value = 0.0821258
standard deviation = sqrt(0.0821258 * (1-0.0821258) / 344660) = 0.000468
margin of error = 1.96 * 0.000468 = 0.00092
lower bound = 0.0821258 - 0.00092 = 0.0812
upper bound = 0.0821258 + 0.00092 = 0.0830
observed = 0.0821824 (given)
The observed value (experiment value) is within the bounds, and therefore this invariant metric passes the sanity check.

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)
Effect Size Tests

Gross conversion

p = (3785 + 3423) / (17293 + 17260) = 0.2086
se = sqrt(0.2086 * (1-0.2086) * (1/17293 + 1/17260)) = 0.00437
d = 3423/17260 - 3785/17293 = -0.02055
lower bound = -0.02055 - 0.00437 = -0.0291
upper bound = -0.02055 + 0.00437 = -0.0120

This metric is statistically significant as the interval does not include zero, and is practically significant as it also includes the practical significance boundary.

Net conversion

p = (2033 + 1945) / (17293 + 17260) = 0.1151
se = sqrt(0.1151 * (1-0.1151) * (1/17293 + 1/17260)) = 0.00343
d = 1945/17260 - 2033/17293 = -0.0048
lower bound = -0.0048 - 0.00343 = -0.0116
upper bound = -0.0048 + 0.00343 = 0.0019

This metric is not statistically significant as it included zero, and therefore not practically significant either.

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

Online calculator is used to perform the sign tests.

Metric	p-value	Statistically significant (< alpha)
Gross Conversion	0.0026	Yes
Net conversion	0.6776	No

Summary

The launch criteria in our case is to test for all of the metrics (Gross conversion and Net conversion) and all of them should be statistically significant

Now the importance of correction is if a test is launched and the metrics shows a significant difference, because it's more likely that one of multiple metrics will be falsely positive as the number of metrics increases.

However, we would only launch if all evaluation metrics must show a significant change and thus, there would be no need to use Bonferroni correction.

Recommendation

I will recommend Not to run this experiment

Following evaluation metrics were used –

Gross Conversion – It is negative and practically significant. This means, we lower our costs by discouraging trial signups that are unlikely to convert

Net conversion – It is statistically and practically non-significant. Additionally, the confidence interval does include the negative number of the practical significance boundary; meaning, it's possible that this number went down by an amount that would hurt the conversion

Basis above analysis, gross conversion result is a good outcome because the cost is lower by discouraging users who're unwilling to commit and unlikely to convert.
However, net conversion results ended up being both statistically & practically non-significant while the confidence interval does include the negative number of the practical significance boundary; meaning, it's possible that this number went down by an amount decrease the revenue.
If we consider the initial hypothesis, it does not increase numbers of paid users, which fails the initial goal of launching this feature and likely to be unacceptable risk to launch.

Follow-Up Experiment

Aim of original experiment - The initial Udacity experiment was focused on acquiring qualified shoppers for better conversion

From the analysis –

- Practically and statistically significant Gross conversion
- A non-significant Net conversion

This means -

- The experiment is successful in reducing cost of futile free trial sign-ups
- It ensures interested candidates sign-up for ‘free trial’
- Enrolled students are confident of being able to give 5 or more hours
- Enrollment post 14 days trial period dipped

Concern Areas –

- User is not confident of value proposition, during 14 days’ free trial

Possible Reasons –

- Benefits of paid courses not promoted properly
- Not able to judge the course material during ‘Free trial’ for 14 days

Follow up experiment -

- During 14 days' free trial, expose review of other students who paid and completed the course. This will create the sense of reassurance among user

Hypothesis – As a user already enrolled for 'free trial', I might not be able to understand the value of paid course within 14 days and might cancel the course enrollment before trial period ends. However, the positive reviews from existing students might influence my decision and thus help me in opting for complete paid course.

Thus this experiment will directly tackle 1st point under 'possible reasons' section and indirectly overcome 2nd reason

unit of diversion: user_ids
This follow-up experiment can use user ids when they sign-up as the unit of diversion. This ensures that a signed-in user is not both in the control and experimental group.

invariant metric: number of user_ids
As the course page changes with half the population only seeing the featured reviews after selecting the start free trial, the number of user that visit the website is unlikely to vary as that page has not been seen yet and should not affect users visiting the page.

evaluation metric: Retention
Since, click and cookie based metrics are decoupled from post enrollment experiments thus retention will be a good evaluation metric as it is User ID based

If the evaluation metric is practically significant and better than the control group at the end of the experiment, we can launch the new feature