Experiment Design

Metric Choice

In this A/B testing, three invariant metrics are identified:

1. Number of cookies
   As the unit of diversion is cookies, both control and experiment groups should have nearly the same number of cookies accessing the course overview page due to even distribution. The metric is measured before the screener popup and thus not affected by the feature, so it will be a good invariant metric.

2. Number of clicks
   The number of unique cookies clicking the “start free trial” button should be evenly distributed across both control and experiment groups, because the screen change happens after the button click and it should not affect the number of clicks in both groups. The metric is measured before the screener popup and thus not affected by the feature, so it will be a good invariant metric.

3. Click-through-probability
   As both number of cookies and number of clicks are invariant metrics, the click-through-probability should be invariant metric too because it is by definition the number of clicks over the number of cookies. It is also derived from metrics collected in the pre-intervention stage so that makes this a good invariant metric.

The following evaluation metrics are identified:

1. Gross conversion
   The hypothesis of the experiment is that the screen popup will reduce the number of frustrated students who left the trial because they do not have enough time. Thus we expect that the number of students to complete the checkout and enroll in the free trial should be reduced if we launch the feature. Gross conversion is the number of user-ids to enroll and complete payment over the number of clicks, so it is exactly the metrics we want to evaluate. It is expected that gross conversion rate should decrease to launch the new feature.

2. Retention
   It is the number of user-ids remain enrolled past the 14 days boundary divided by the number of user-ids to complete checkout. Users in experiment group may think twice before they enrol the course due to the screener. As the screener popup may reduce the number of enrolment, it is not an invariant metric. Thus it is expected that the retention

rate of experiment group may be higher than the rate of control group, and it makes retention a good evaluation metric.

3. Net conversion
   In the experiment group, the number of user-ids passing the 14-day trial period after they click “start free trial” button may or may not be affected by the reminder message. As the number of users passing the 14-day trial is recorded in the post-intervention stage, it is not good invariant metric because we cannot tell whether the popup will increase or decrease the number of user-id passing the 14-day trial. It should be made an evaluation metric so we could know whether the new feature will affect the number of students who complete the course. Ideally this metric should remain unchanged, because the hypothesis is that the feature will only reduce the number of frustrated students who left the 14-day trial period.

The number of user-id is not used as both invariant and evaluation metric. It is not a good invariant metric because the number of enrolment should be affected by the new screener feature. It is not a good evaluation metric because the numbers of user-id of the control and experiment group is not a normalized value (like gross conversion) and it is difficult to compare this unevenly distributed counts.

Given the above evaluation metrics, the launching criteria will be:

- Gross conversion decreases. This is the result we expect the new feature can bring
- Retention or Net conversion remain unchanged or even increase. However, to launch the change, these metrics must not decrease.

Measuring Standard Deviation

Evaluation Metrics	**Standard Deviation**
Gross Conversion	0.0202
Retention	0.0549
Net Conversion	0.0156

Gross conversion

It 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. As both unit of analysis and unit of diversion is cookies, I expect the analytic estimate would be comparable to the empirical variability.

Retention

The standard deviation of retention will likely be different from the empirical estimate, because the unit of analysis of retention is user-id and that is different from the unit of diversion.

Net conversion
Like the gross conversion, the unit of analysis of this metric is cookies and it is the same as the unit of diversion. Therefore the analytic estimate should be comparable to the empirical variability.

Sizing

Number of Samples vs. Power

I will not use Bonferroni correction because the evaluation metrics are highly correlated to each other. By using an alpha of 0.05 and a beta of 0.2, the number of pageviews needed for each evaluation metric is calculated accordingly. The sample size needed for each metric is calculated using the online calculator.

Gross Conversion

- Baseline conversion rate = 20.625%
- d_{min} = 1%
- Sample size needed = 25835
- Number of pageviews needed = (25835 / 0.08) * 2 = 645,875

Retention

- Baseline conversion rate = 53%
- d_{min} = 1%
- Sample size needed = 39115
- Number of pageviews needed = (39115 / (660 / 40000)) * 2 = 4,741,212

Net Conversion

- Baseline conversion rate = 10.93125%
- d_{min} = 0.75%
- Sample size needed = 27413
- Number of pageviews needed = (27413 / 0.08) * 2 = 685,325

The number of pageviews needed is 4,741,212, which is a very large number. The duration to generate the traffic is unrealistic for a A/B testing (see next section). Thus retention will not be used as evaluation metric. The number of pageviews needed for this experiment will be 685,325.

Duration vs. Exposure

It is expected that risk of running the experiment is very low. There will be no sensitive data being collected because it only gives user an option to early discontinue the process. For the users in the experiment group, there will not be any kind of damages. Even we later withdraw the change from the experiment, there will be no side-effect for the experiment subjects.

Provided that there is no other experiment running at the same period, I think a 100% exposure is not risky to the company. The following shows the number of days needed for 100% exposure, given that daily pageviews is 40000.

Evaluation Metrics	Number of Pageviews needed
Gross Conversion	(645875 / 40000) = 16.15
Retention	(4741212 / 40000) = 118.53
Net Conversion	(685325 / 40000) = 17.13

As the retention metric needs 119 days to collect the needed pageviews, it is not suitable for evaluation metric in A/B testing. For 100% exposure, it will take 18 days to complete the testing, which is a suitable duration for A/B testing.

Experiment Analysis

Sanity Checks

Number of cookies
# cookies (control) = 345543
# cookies (experiment) = 344660
Standard deviation = sqrt(0.5*0.5 / (345543 + 344660)) = 0.0006018
Margin of error = 1.96 * 0.0006018 = 0.00118
Bounds = [0.5 - 0.00118 , 0.5 + 0.00118] = [0.4988 , 0.5012]
Observed value = 345543 / (345543 + 344660) = 0.5006

As the observed value is within the bounds, this metric passes the sanity test.

Number of clicks
# clicks (control) = 28378
# clicks (experiment) = 28325
Standard deviation = sqrt(0.5*0.5 / (28378 + 28325)) = 0.0021
Margin of error = 1.96 * 0.0021 = 0.0041
Bounds = [0.5 - 0.0041 , 0.5 + 0.0041] = [0.4959, 0.5041]

Observed value = 28378 / (28378 + 28325) = 0.5005

As the observed value is within the bounds, this metric passes the sanity test.

Click-through-probability
Click-through-probability (control) = 28378 / 345543 = 0.0821
Standard deviation = sqrt(0.0821 * (1 - 0.0821) / 344660 = 0.000468
Margin of error = 1.96 * 0.000468 = 0.0009
Bounds = [0.0821 - 0.0009 , 0.0821 + 0.0009] = [0.0812 , 0.0830]
Observed value = (28325 / 344660) = 0.0822

As the observed value is within the bounds, this metric passes the sanity test.

Result Analysis

Effect Size Tests

Gross Conversion

enrol_control	3785
clicks_control	17293
enrol_exp	3423
clicks_exp	17260
Gross conversion (p) = (enrol_control + enrol_exp) / (clicks_control + clicks_exp)	0.2189
Standard Deviation = sqrt(p * (1-p) * (1/clicks_control + 1/clicks_exp))	0.0044
d = (enrol_exp/clicks_exp) - (enrol_control/clicks_control)	-0.0206

Bound = [d - 1.96*0.0044 , d + 1.96*0.0044] = [-0.0291 , -0.012]

As the bound does not cover zero, it is statistically significant. It is also practically significant because the bound does not cover the practical significance (d_{min}) of 0.01.

Net Conversion

payment_control	2033
clicks_control	17293
payment_exp	1945

clicks_exp	17260
Gross conversion (p) = (payment_control + payment_exp) / (clicks_control + clicks_exp)	0.1151
Standard Deviation = sqrt(p * (1-p) * (1/clicks_control + 1/clicks_exp))	0.0034
d = (payment_exp/clicks_exp) - (payment_control/clicks_control)	-0.0049

Bound = [d - 1.96*0.0034 , d + 1.96*0.0034] = [-0.0116 , 0.0019]

As the bound includes zero, it is neither statistically significant and practically significant.

Sign Tests

The click/enrolment date starts from Oct 11 to Nov 2, which is totally 23 days.

The sign test for gross conversion shows that 4 out of 23 days have positive increase. It means a two-tailed p-value of 0.0026, which is smaller than the alpha (0.05). Thus the gross conversion metric passes the sign test and is statistically significant.

The sign test for net conversion shows that 10 out of 23 days have positive increase. It means a two-tailed p-value of 0.6776, which is much larger than the alpha (0.05). Thus the net conversion metric fails the sign test and is not statistically significant.

Summary

I have not used the Bonferroni correction. This correction has a characteristics that if the metrics are highly correlated to each other, the correction will be too conservative in order to pass. In this test the gross conversion and net conversion is highly correlated so the correction may not be appropriate. More importantly, in this test we have two metrics to consider and ideally we want BOTH metrics be statistically significant in order to launch. Using Bonferroni correction will control the false positives in the cost of power, and it is against our purpose because we want ALL metrics meeting the expectation.

The results of effect size hypothesis tests and the sign tests align with each other. The gross conversion metric is both statistically and practically significant, and the net conversion metric is not statistically significant.

The experiment shows that the change will significantly decrease the gross conversion rate and keep the net conversion rate unchanged.

Recommendation

I would recommend not to launch the change. The test result shows that it can decrease the gross conversion, which matches our expectation because we hope that this change can reduce the number of frustrated students who drop out within the 14-days trial and save more coaching resources for other students. However, when we look at the net conversion rate, we notice that the CI lower bound (-0.0116) is smaller than the negative practical significance (-0.0075). It means that it is possible for the net conversion rate to drop after we launch the feature, and it will hurt the business bottom line. We can only launch if the net conversion CI falls within the practical significance.

Follow-Up Experiment

When we perform A/B testing for the “free trial screener” feature, we notice that the retention rate cannot be measured due to the lengthy duration. A follow-up experiment could try to increase the retention rate as we ultimately want more students to pass the 14-days trial.

I would suggest a “keep the momentum” experiment, which will encourage the enrolled students by using positive wordings for their hard work during the 14-days trial. The purpose of this experiment is to test whether we could influence more students to pass the 14-days trial during the trial period. With the assumption that the system can keep track of the learning hours of each student, we could display a “keep the momentum” notice to students who spend more than X hours a week after they login. We could display the notice on day 7, i.e. half of the 14-days trial. The number of hours will be something less than 5 hours (our recommended effort) so that we could encourage the students with the potential to continue. The number X and the notice day should be discussed with business user, and the assumption about the learning hour records needs confirmation from engineering team too. One hypothesis is that we may not select a too low value for X, because it may actually discourage students. Note that the overall goal of “keep the momentum” experiment is to increase the retention rate, which is quite different from the “free trial screener” experiment with goal to save more resources for potential students.

The evaluation metrics used for this experiment will be retention and net conversion. The invariant metrics will be the number of user-ids which completed enrollment because we could evenly distributed it across control and experiment groups as the unit of diversion. Other metrics like number of cookies visiting course page and number of clicks on start free trial are not good invariant metrics because those are pre-enrollment metrics and this experiment is a post-enrollment experiment. As the main metric we want to test is retention, the unit of diversion will be user-id. The hypothesis is that the change will increase the retention rate with a practical significance (e.g. 0.01) as agreed by stakeholders.