DAND P7: A/B Testing for Udacity Free Trial Screener

Experiment Overview

In this experiment, Udacity tested a change to the process of free trail enrollment for the purpose of reducing the number of drop-offs during the free trial due to lack of time commitment. The free trial screener window popped up after a user clicked on “Start Free Trial” and asked for devoted time allocation. If the potential student was not able to commit more than 5 hours of study time, a message would appear to warn the student that it was not good practice and gave the option of free access to course materials instead.

Experiment Design

The unit of diversion in this experiment is the cookie. Once a student enrolls in the free trial, the experiment track the user ids. The same user id cannot enroll twice.

Metric Choice

Table 1 shows and explains the definition and properties of 7 potential metrics for the experiment. For each metric, whether it is measured before or after the appearance of screener will affect the expected change due to the screener, as difference between control and experiment groups. Generally speaking, normalized changes are good choice for evaluation metrics.

Table 1 Potential choices of metrics

	Definition	Measured before or after the screener?	Expected Change due to the screener?	Choice as Invariant or Evaluation Metric?	Comments (Reason for choice)
Number of cookies	number of unique cookies to view the course overview page	Before	None	Invariant	Should be evenly distributed among control and experiment group. Should be verified by sanity check.
Number of user-ids	number of users who enroll in the free trial	After	Decrease	Neither	Not normalized. Fluctuates from day to day. Not a good choice as metric.
Number of clicks	number of unique cookies to click the "Start free trial" button	Before	None	Invariant	Should be evenly distributed among control and experiment group. Should be verified by sanity check.
Click-through-probability	number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page	Before	None	Invariant	Should be similar for control and experiment group. Should be verified by sanity check.

Gross conversion	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	After	Decrease	Evaluation Metric	Normalized change due to screener. It should decrease due to the warning imposed to the users.
Retention	number of user-ids to remain enrolled past the 14-day boundary divided by number of user-ids to complete checkout	After	Increase	Evaluation Metric	Normalized change due to the screener. The screener reduced the drop-offs during the free trial.
Net conversion	number of user-ids to remain enrolled past the 14-day boundary divided by the number of unique cookies to click the "Start free trial" button	After	No change or increase	Evaluation Metric	Normalized change. It may or may not be affected by the change. It should not decrease.

In summary, the following were chosen as the invariants:

- Number of cookies
- Number of clicks
- Click-through-probability

The following were chosen as the evaluation metrics:

- Gross conversion
- Net conversion

Here it is necessary to note that 'Retention' is also a good choice for evaluation metric, however, it was dropped later for this project due to the large sizing it required. It will be shown in a section 'Sizing' that this metric would require a long duration of experiment.

Measuring Variability

The analytical estimation of standard deviation of evaluation metrics are shown in Table 3. The n value used in the calculation was obtained from scaling down from baseline value to the sample size of 5000 (shown in Table 2)

Table 2 Size scaling for sample size of 5000

	Baseline	Sample
Unique cookies to view page per day	40000	5000
Unique cookies to click "Start free trial" per day	3200	400
Enrollments per day	660	82.5

Table 3 Calculation of standard deviation of potential evaluation metrics

	Baseline Probability	SD=sqrt(p*(1-p))	n*	S.E.=SD/sqrt(n)
Gross conversion	0.2063	0.4046	400	**0.0202**
Retention	0.5300	0.4991	82.5	**0.0549**
Net conversion	0.1093	0.3120	400	**0.0156**

It is noted the SD in Table 3 is the standard deviation of the individual values. Standard error (SE) instead should be used to measure the variability of the evaluation metrics. In fact, SE is the standard deviation of the sample means.

For 'Gross conversion' and 'Net conversion', the unit of analysis is 'cookies', which is the same as unit of diversion. Therefore, we could expect binomial distribution on these metrics. For these two metrics, the analytical estimation of variability is expected to be fairly accurate.

For 'Retention', since the unit of analysis is 'user-id', the analytical estimation would not be accurate. If it is chosen as an evaluation metric, it is important to estimate the standard deviation empirically. Since 'Retention' was dropped eventually, the empirical calculation is not necessary here in this project.

Sizing

Number of Samples vs. Power

Due to the dependency of metrics, Bonferroni correction was not used for this analysis. An online sizing estimator has been used to calculate the required size of experiment. With:

$$\propto = 0.05$$

$$\beta = 0.2$$

The calculation of sample sizes is shown in Table 4.

Table 4 Sizing of Experiment

	Baseline Conv. Rate	d_{min}	Sample Size Required	# Pageviews Required*	Total pageviews Required (2 groups)
Gross conversion	0.2063	0.01	25835	322938	645876
Retention	0.5300	0.01	39115	2370606	4741212
Net conversion	0.1093	0.0075	27413	342663	**685325**

*Calculated based on the same principal in Table 2.

It is noted in the table, the calculation of required size for 'Retention' are greyed out since it would require an enormous amount of data, which became impractical to run the experiment. Therefore, it was dropped from the evaluation metrics. In the final decision, 685325 had been selected as it was the larger one between 'Gross conversion' and 'Net conversion'.

Duration vs. Exposure

The risks exposed on the users by the experiment can be summarized as:

1. Psychological: the reminded commitment of dedicated study time may be discouraging to some users. This may make then avoid signing up for any programs.

2. Physical: user need to go through one extra step before they could get to the checkout page. This is one more click.
3. Financial: The user might potentially lose the opportunity to learn new things thus a better financial future.

While the first risk is expected to be a valid concern, the second and third risks are minimal. If majority of the users are not exposed to the risks, it would be an experiment worth running.

As the daily pageview of 40000, 75% of the traffic would be diverted to the experiment. This means only about one third (37.5%) of the traffic would experience the experimental change, so it is not too risky to run. This setting would require the number of days for the experiment as:

$$685326 \div 0.75 \div 40000 = 22.84 \approx 23\ days$$

It is noted that if 'Retention' had been chosen as an evaluation metrics, it would require the experiment to last 158 days or 23 weeks. This would have been impractical for such as small change to be implemented.

Experiment Analysis

The experiment had been run for 37 days. There are 37 days of data on pageview and click. For data on enrollment and payments, there are only for 23 days, since a 14-day period were needed to find out whether a user kept staying in the program after free trial ended.

Sanity Check

Sanity check should confirm that the metric selected as invariant should be similar among the controlled group and experiment group.

For each of the invariants, 95% confidence interval had been calculated and compared with the observed value from the experiment. For 'Number of cookies' and 'Number of clicks', the intervals were calculated around 0.5 probability for controlled group. For 'Click-through-probability', the confidence interval had been calculated for difference between the two groups around zero. The calculation and sanity check results are shown in Table 5.

Table 5 Sanity Check Results

Metric	Expected value	95% Confidence Interval	Observed Value	Pass Sanity Check?
Number of Cookies (ratio of control group)	0.5	[0.4988, 0.5012]	0.5006	Yes
Number of Clicks (ratio of control group)	0.5	[0.4959, 0.5041]	0.5005	Yes
Click-through-probability (difference between groups)	0	[-0.0013, 0.0013]	0.0001	Yes

It is shown that all three invariants passed the sanity check. This mean the results from the experiments are most likely valid and can be further analyzed.

Result Analysis

Effect Size Tests

With 95% confidence level, the confidence interval has been calculated around the difference of evaluation metrics between the controlled and experiment groups. Bonferroni correction was not used. Therefore, the confidence interval of each metrics had been examined at 95% confidence level. The results are shown in

Table 6.

Table 6 Results from Effect Size Tests

Metric	95% Confidence Interval	Statistically Significant?	d_{min} (Practical Significance Level)	Practical Significant?
Gross Conversion	[-0.029, -0.0120]	Yes	0.01	Yes
Net Conversion	[-0.0116, 0.0019]	No	0.0075	No

Sign Tests

A sign test had been performed on day-by-day data for the 23 days. Again, due to dependency of the metrics, Bonferroni correction had not been used. An online calculator has been used to calculate the overall probabilities. The results of sign test are shown in Table 7.

Table 7 Results from Sign Tests

Metric	Number of Days with Predicted Change	Number of Days of Experiement	Assumed Probability of Predicted Change	Overall Prob. of Changes	Statistical Significant?
Gross Conversion	19	23	0.5	0.0026	Yes
Net Conversion	10	23	0.5	0.6776	No

Summary

Bonferroni correction was not used in any of the analysis for this project. The main reason was that the two evaluation metrics were dependent, and the use of the correction may be too conservative.

From the above results, it can be seen that effect size tests and sign tests produce the similar results, which is that 'Gross Conversion' has statistical significant change, while 'Net Conversion' did not show significant change. Therefore, there is no discrepancy between the two tests.

Recommendation

The recommendation based on the experiment results is **not to launch** the change. Instead, additional testing needs to be run. This is based on the following:

1. 'Net Conversion' was not significant enough to show the final number of student who remained in the program would not be affected by this change.
2. Significant decrease of 'Gross Conversion' means that number of student enrolled in the free trial had been decreased. Without confirming 'Net Conversion' not being affected or significant increase of 'Retention', this could be a dangerous move. It could potentially mean some

students who could have sign on and finish the program were scared away by the screener which requested for time commitment.

3. In order to proof the decrease of drop-offs during free trial, the metric 'Retention' needs to be used. This would require a longer duration for the experiment.

Follow-up Experiment

In order to reduce the number of students who drop out during the first 2 weeks, students need to have the right expectation of the course, which includes time commitment, prerequisite and study outcome. An experiment I would run is to give student an **optional** screening test. The test should be sufficient to identify the student's potential obstacles in following the program, and it should be simple and short enough that would not impose too much burden to the student. If the student pass the test, he/she could skip some part of the program or be given some credit.

Experiment Setup: An optional screening test is popped up after clicking 'Start Free Trial' for the experiment group. The unit of diversion is the 'cookie'.

Null Hypothesis: The drop-off rate during the first 14 days in the experiment group is not lower than that in the controlled group.

Alternative Hypothesis: The drop-off rate in the experiment group overall is significant lower than controlled group.

Choice of Invariants: Number of cookies; Number of Clicks; Click-through-probability;

Choice of Evaluation Metrics:

- Gross conversion: It should be expected to decrease due to the student were given the clearer expectation.
- Net conversion: This should not decrease as the change should not have negative impact on those students who were ready to commit to the course/program.
- Retention: This is the opposite of drop-off rate. So it would be expected to increase.
- Tested Retention (The number of user-ids in the experiment group who took the test and remained enrolled after 14 days divided by number of cookies who took the screening test): This should increase as the students demonstrate their commitment by taking the test.
- Non-tested Retention (The number of user-ids who didn't take the test remained enrolled after 14 days divided by the number of cookies who didn't take the screening test): This should not decrease, as the student would be confident enough to skip the test.

References:

[1] Sample Size Calculator http://www.evanmiller.org/ab-testing/sample-size.html

[2] Sign test http://graphpad.com/quickcalcs/binomial1/