A/B Testing for Udacity Free Trial Screener

By: W. Alexander Jenkins

Experiment Design

Metric Choice

- Invariant metrics: Number of cookies, Number of clicks
- Evaluation metrics: Gross conversion, Net conversion

Figure 1 shows a high-level view of the customer funnel; it will help to draw boundaries between events in the free trial enrollment process. In between the click “start free trial” and enrollment stages, the experiment group will be presented with the free trial screener.

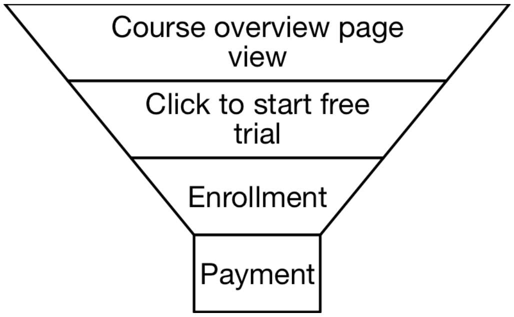

Figure 1. Customer funnel for Udacity courses

Reasoning for Metric Choice

Number of cookies (Number of unique cookies to view the course overview page):
The unit of diversion is a cookie so we expect the number of cookies should be comparable for the experiment and control. This metric is independent from the experiment; the course overview page view occurs before the change in the experiment. Its independence also makes it a poor choice for an evaluation metric.

Number of user-ids (Number of users who enroll in the free trial):
This metric would not make a good choice for an invariant metric; we can expect this number to vary across the experiment and control groups. Enrolling in the free trial occurs after the change being tested. We want to reduce the number of students that leave the free trial. This metric could be used as an evaluation metric since it tracks the number of users to enroll in the free trial. It could test the number of students who continue past the free trial. However, it is not the best choice because it is not normalized for “Start free trial” clicks.

Number of clicks (Number of unique cookies to click the “Start free trial” button):
This metric is independent of the change; it occurs right before the change. We expect this number to be the same across the experiment and the control; we can ensure the sizes of each group are comparable. It will be used as an invariant metric. This would make a poor choice for an evaluation metric due to its independence from the change.

Click-through-probability (Ratio of unique cookies who click “Start free trial” to unique cookies that view the course overview page):
The click-through-probability would not make a good choice for an evaluation metric; the clicks occur before the change being evaluated. The metric tracks the number of clicks, a metric we need to track, while normalizing for page views. It could be used as an invariant metric, but this experiment will use the number of clicks instead.

Gross conversion (Ratio of user-ids that complete checkout and enroll in the free trial to unique cookies that click the “Start free trial” button):
This metric would not make a good choice for an invariant metric; the event is dependent upon the change. We expect the number of user-ids to differ between the experiment and the control groups. This experiment uses gross conversion as an evaluation metric. This test will look for the gross conversion of the control group to be greater than the gross conversion of the experiment group. This could show that we’ve diverted users who may not have enough time to devote to the course.

Retention (Ratio of user-ids that remain enrolled past the free trial period and make at least one payment to user-ids that complete checkout and enroll in the free trial):
This metric would not make a good choice for an invariant metric; this metric would allow us to test the hypothesis. It would make a good choice for an evaluation metric. This experiment will not use it since retention can be calculated using the other two evaluation metrics.

Net Conversion (Ratio of user-ids that remain enrolled past the free trial period and make at least one payment to unique cookies that click the “Start free trial” button):
This metric cannot be used as an invariant metric; the event is dependent upon the change. This experiment will use net conversion as an evaluation metric and evaluate whether the free trial screener change helps improve the overall experience. We will look for no decrease between the experiment and control; this would indicate we haven’t reduced the number of students that continue past the free trial.

Measuring Standard Deviation

TABLE 1 – EVALUATION METRIC STANDARD DEVIATIONS

Evaluation Metric	Standard Deviation
Gross conversion	0.0202
Net conversion	0.0156
Retention	0.0549

The standard deviations in Table 1 were computed based on 5000 page views from the baseline of 40000 page views and 3200 clicks. The unit of analysis and the unit of diversion are the cookie for both the gross conversion and the net conversion. The analytical estimate would be comparable to the empirical variability for the net conversion and gross conversion. The unit of analysis for retention is user-id. We likely have an under estimate for the retention since the standard error is proportional to the sample size.

Sizing

Number of Samples vs. Power

The Bonferroni correction will not be used during the analysis phase. The number of page views needed to power the experiment is 685,325. 4,741,212 page views would be needed to power the experiment if retention was included as an evaluation metric. This total is not feasible for this experiment, as it would take too long to run. It would take 18 weeks to run the even if 90% of the traffic was diverted to this experiment.

Duration vs. Exposure

If we divert 60% of the traffic to this experiment, it would take 29 days to run. The change does not exceed the minimal risk threshold; it does not raise undue physical, psychological, or economic harm to the user. The experiment does not invade the student's privacy, and the data being collected is not sensitive. All traffic could be diverted to this experiment. However in case there any bugs in the change or process, we will not divert everyone to the experiment.

Experiment Analysis

Sanity Checks

Invariant Metrics

Confidence interval on the fraction of cookies I expect to observe in the control group:
95% CI = (0.4988,0.5012). I observed 0.5006. This sanity check passes.

Confidence interval on the fraction of clicks I expect to observe in the control group:
95% CI = (0.4959,0.5041). I observed 0.5005. This sanity check passes.

Result Analysis

Effect Size Tests

Confidence interval around the difference between the gross conversion of the experiment and control groups: 95% CI = (-0.0291, -0.0120)
The gross conversion is statistically significant (does not contain 0) and practically significant (does not contain d_{min} = 0.01).

Confidence interval around the difference between the net conversion of the experiment and control groups: 95% CI = (-0.0116, 0.0019)
The net conversion is not statistically significant (interval contains 0) or practically significant (contains $-d_{min}$ = -0.0075).

Sign Tests

Gross conversion sign test p-value = 0.0026 < alpha = 0.05. Gross conversion is, in fact, statistically significant.
Net conversion sign test p-value = 0.6776 ≥ alpha = 0.05. Net conversion is definitely not statistically significant.

Summary

Bonferroni correction was not used because it is too conservative. The net conversion and the gross conversion are correlated. In this experiment these metrics both need to match our expectations and show significance in order for Udacity to launch the change. Gross conversion should show a decrease while net conversion does not decrease. The Bonferroni correction would be used if we needed any one metric to match our expectations. We need all evaluation metrics to match our expectations.

Recommendation

Udacity should not launch the change. The gross conversion was negative and statistically and practically significant. The change successfully reduces the number of students enrolling in the free trial but showing the free trial screener. This reduction in students could expand the reach of coaches' capacity to offer support.

However, the net conversion was neither statistically or practically significant. Enrollees are still cancelling their free trials. The warning shown to the students in the experiment group did not significantly increase the proportion of students who stick around after the free trial and make a payment. The confidence interval (95% CI: [-0.0116, 0.0019]) contains the negative of the practically significant value -0.0075. We reject the hypothesis; the free trial screener might have reduced the net conversion (and possibly revenue) by an amount that matters to Udacity.

Follow-Up Experiment

A follow up experiment Udacity could run is when a student logs in to the classroom homepage, a popup dialog box appears. The popup would show words of encouragement or a motivational quote followed by a link to Udacity coaching support. The popup itself would be encouragement for the student to push forward with the course and to seek help if needed. The null hypothesis is that adding the popup dialog box to the login process would not significantly increase retention.

The unit of diversion is the user-id. Users are assigned user-ids after enrollment in the free trial. Each time the user logs in after enrollment, they are shown the dialog popup. The invariant metric will be the number of user-ids; we can expect this number to be comparable across the experiment and the control groups. We can make sure to show the popup only once per user-id per day. Retention, or the number of user-ids that continue past the free trial period divided by the number of user-ids in the free trial, will be used as an evaluation metric. A practically and statistically significant increase will show a decrease in the number of students who cancel early. Then Udacity could launch the change.

References

http://www.evanmiller.org/ab-testing/sample-size.html
http://graphpad.com/quickcalcs/binomial1/
https://discussions.udacity.com/c/nd002-p7-design-an-a-b-test