Experiment: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead. This screenshot shows what the experiment looks like.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

Experiment Design

Metric Choice

The following table shows the metric choices in this A/B test.

Type	Metric	Reason
Invariant Metrics	Number of cookies: that is, number of unique cookies to view the course overview page (d_{min}=3000)	The trial screen was triggered after students clicking the ‘‘Start free trial’ button, which is after students viewing the course overview page. Therefore, whether the question pops after will not affect the number cookies viewing the course overview page.
	Number of clicks: that is, number of unique cookies to click the ‘Start free trial’ button. (d_{min}=240)	Since the trial screen was triggered after students after students clicking the ‘‘Start free trial’ button, the number of clicks on the

		button should not change with the experiment.
	Click-through-primality: that is, the number of unique cookies to click the ‘Start free trial’ button divided number of unique cookies to view the course overview page. (d_{min}=0.01)	As illustrated before, both of the divisor and dividend are recorded before the screener. Therefore, the fraction could be used as an invariant metric.
Evaluation metrics **Measure the impact of the screener on free trial completion and payments.**	Retention: that is, number of user-ids to remain enrolled past the 14-day boundary(and thus make at least one payment) divided by number of user-ids to complete the checkout. (d_{min}=0.01)	Since the experiment is designed to test if the free trial screener would help increase the percentage of students enrolled after checking out, the retention would be a good evaluation metric. However, this metric is dropped in the final experement, which will be describled in details i later.
	Net conversion: that is, 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. (d_{min}= 0.0075)	For the similar reason, net conversion can measure the impact on free trial completion.
Evaluation metrics **Measure the impact of the screener on enrollment**	Gross conversion: That is, 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. (d_{min}= 0.01)	Gross conversion tries to measure the percentage of users that check out after clicking the button.
Metrics not useful in this case	Number of user-ids: That is, number of users who enroll in the free trial.	Enrolling happening after the trial screening and cannot be used as an invariant metric. Also it is not ideal for the evaluation metric since number of user-ids is a raw count. A raw count cannot adjust to different sized experiment and control groups and therefore cannot account for the real change of the enrollment rate.

The following table shows the launch criteria in this A/B test.

What to measure	Metric	Launch Criteria
The impact of the screener on free trial completion and payments.	Net conversion	There is **NO** statistically significant **decrease** in net conversion between control group and experiment group.
The impact of the screener on enrollment	Gross conversion	There is a statistically significant **decrease** in net conversion from control group to experiment group.

Measuring Standard Deviation

Analytically computed standard deviation for evaluation metrics:

Retention	0.0549
Net conversion	0.0156
Gross conversion	0.0202

The analytic estimates are accurate?

- Indepedence

When the unit of diversion and unit of analysis are the same, the analytically computed variability is likely to be very close to the empirically computed variability.

In our case:

	Unit Of Analysis	**Unit Of Diversion**
Retention	user-id	cookie
Net conversion	cookies	cookie
Gross conversion	cookies	cookie

For net conversion and gross conversion, two units are consistent. Therefore, the analytic variance can be used.

For retention, we are using the cookie-based diversion but the unit of analysis is user-id. The independence assumption is not valid since we are diverting a subgroup of the unit of analysis. So we want to shift to empirically computed variability.

- Normality

Since both np and n(1-p) are bigger than 5. A normal distribution assumption would be reasonable. Since all the metrics follow a binomial distribution, they are relatively simple, analytic estimate may be a good option here. The analytic estimate should be comparable to the empirical variability.

If we were analyze more complicated metrics, the distribution can be very weird and we may want to shift to an empirical estimate.

Sizing

Number of Samples vs. Power

- An alpha of 0.05 and a beta of 0.2 are used
- Bonferroni correction is NOT used
- Online calculator is used: http://www.evanmiller.org/ab-testing/sample-size.html
- Retention: d_{min}=0.01
- Net Convention:d_{min}= 0.0075
- Gross conversion: d_{min}= 0.01

The pageviews total (across both groups) needed to collect to adequately power the experiment are:

Sample size	Retention	Net conversion	Gross conversion
Per variation	39115	27411	25835
In total	39115 * 2 * 5000/82.5 = 4,741,212	27411 * 2 * 5000/400 = 685,275	25835 * 2 * 5000/400 = 645,875

Since 4.7 million pageviews would take 119 days, we are going to drop the metric-retention-favoring another metric- net conversion - to measure impact of the screener on free trial completion and payments.

Max(645,875, 685,275)= 685,275
We need 685,275 pageviews.

Duration vs. Exposure

685,275/40,000 = 17.13 $\approx$ 18 (days)
If 100% of the traffic are diverted to this experiments, we would need 18 days.

In this case, we think it is okay to divert 100% of the traffic for the following reason:

- To save time
- There is no indication that other experiment needs to be done during the same period
- This experiment does not sensitive information since it is based on cookie.
- The experiment does not harm students since it is only reminder that the course takes time.

Experiment Analysis

Sanity Checks

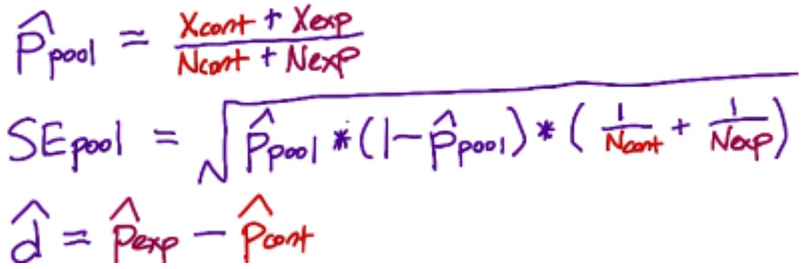
For each of the evaluation metrics, a 95% confidence interval around the difference between the experiment and control groups.

First, number of cookies and clicks are evaluated.

	Pageviews	Clicks
Control	345543	28378
Experiment	344660	28325
Total	690203	56703
SE	sqrt(0.5*0.5/690203)=0.0006	sqrt(0.5*0.5/56703)=0.0021
Margin of Err	1.96*sd=0.0012	1.96*sd=0.0041
Upper Bound	0.5011796079	0.5041155043
Lower Bound	0.4988203921	0.4958844957
Observed (control)	0.5006396669	0.5004673474

The observed value is within the confidence level. The difference between control and experiment group is not statistically significant with an alpha of 0.05 when it comes to number of cookies or clicks.

Then we look at the click-through-primality:



$$\hat{P}_{pool} = \frac{X_{cont} + X_{exp}}{N_{cont} + N_{exp}}$$

$$SE_{pool} = \sqrt{\hat{P}_{pool} * (1 - \hat{P}_{pool}) * \left(\frac{1}{N_{cont}} + \frac{1}{N_{exp}}\right)}$$

$$\hat{d} = \hat{P}_{exp} - \hat{P}_{cont}$$

P hat pool = 0.08215
SE pool = 0.00066
d hat = 0.00006 < Margin of Err = 0.00130
CI: [-0.0013,0.0013]

The observed difference between proportion is smaller than the margin of error. The difference between control and experiment group is not statistically significant with an alpha of 0.05.

The sanity check is passed.

Result Analysis

Effect Size Tests

For each of the evaluation metrics, alpha = 0.05.

	Gross conversion	**Net conversion**
Diff (exp-con)	-0.02055	-0.00487
p hat pool	0.20861	0.11513
1/Ncon+1/Nexp	0.00012	0.00012
SE pool	0.00437	0.00343
Margin of Err	0.00857	0.00673
Upper Bound	-0.0120	0.0019
Lower Bound	-0.0291	-0.0116
z score	4.70183	1.41920
P value	<0.0001	0.15580

The difference for gross conversion is statistically significant. Also, it is practically significant since $0.0206>d_{min}=0.01$

The difference for net conversion is not statistically significant nor practically significant. (d_{min}= 0.0075)

Sign Tests

Con-Exp	Gross conversion	Net conversion
Sat, Oct 11	0.04198972165	0.05232960308
Sun, Oct 12	0.04093276535	-0.02606477355
Mon, Oct 13	0.01969122252	0.01514393521
Tue, Oct 14	0.01973467251	0.01435262059
Wed, Oct 15	0.02647389946	-0.0365172089
Thu, Oct 16	0.003973638601	0.02222431244
Fri, Oct 17	0.03236665295	0.04519402166

Sat, Oct 18	0.02987885377	0.01566746913
Sun, Oct 19	0.01741389083	-0.0236427775
Mon, Oct 20	0.01373065392	-0.001293790347
Tue, Oct 21	0.06055763809	0.03893134051
Wed, Oct 22	0.03351717275	0.02239444132
Thu, Oct 23	0.000946290961	-0.0217084768
Fri, Oct 24	0.0485587777	0.04641415875
Sat, Oct 25	0.06486775302	0.06416255119
Sun, Oct 26	0.006621909164	0.001149509624
Mon, Oct 27	0.03071828076	0.00902785254
Tue, Oct 28	-0.01086956522	-0.03940217391
Wed, Oct 29	-0.01125540481	-0.0156760409
Thu, Oct 30	-0.05682022337	0.01014686948
Fri, Oct 31	-0.005618638814	-0.02730062072
Sat, Nov 1	0.02475834311	-0.007321015434
Sun, Nov 2	0.04587708179	-0.04558409669
Con<Exp?	4	10
Total	23	23

Using the online calculator, we get the following sign test results:

	Gross conversion	Net conversion
Number of "successes"	19	10
Number of trials	23	23
The P value	0.0026	0.6776
Explanation	the chance of observing either 19 or more successes, or 4 or fewer successes, in 23 trials	the chance of observing either 10 or fewer successes, or 13 or more successes, in 23 trials.

The sign test does agree with the confidence interval for the difference.
The difference for gross conversion is statistically significant.
The difference for net conversion is not statistically significant.

Summary

I did not use Bonferroni correction because it controls for false positives at the expense of increased false negatives.

Our launch criteria is meeting both criteria: retaining the hypothesis for net conversion AND rejecting the hypothesis for gross conversion. When both metrics must be satisfied to trigger launch, a false negative for gross conversion or a false positive for net conversion can govern the decision.

Therefore, we do not want to control Type I error (incorrectly rejecting) at the risk of increasing Type II error(incorrectly retaining).

Recommendation

I do not recommend this change.

There is statistically significant and practically significant difference in cross conversion. The decrease in gross conversion due to the screener is statistically significant, which is up to our expectation.

However, the negative of the net conversion practical significance boundary (-0.0075) is within the net conversion confidence interval (-0.0116,0.0019). It means the actual change of net conversion may be bigger than -0.075. We cannot risk the potential of revenue drop.

Follow-Up Experiment

To find a potential way to reduce the number of frustrated students who cancel early in the course, I would like to run an experiment on extending the trial period by another week to those students who want to cancel early. In this way, they may have more time to overcome the difficulty they ran into in the course.

Hypothesis: By popping up the screener indicating the course can be extended for another week for free, there will be fewer early cancellation.

Metric:
Invariant metrics: same as the original experiment.
Evaluation metrics: net conversion: that is, number of user-ids to make at least one payment divided by number of unique cookies to click on the ‘Free Trial’ button. (dmin=0.0075)

The unit of diversion and analysis is cookie. Choosing cookie as the unit of diversion and analysis will decrease the sample size based on the previous study.