Experiment Overview

Free Trial Screener

Udacity courses have two options on the home page: "start free trial" and "access course materials":

- If the student clicks "start free trial", he or she is prompted to enter credit card information before being enrolled in a free trial for the paid version of the course. The free trial period lasts for 14 days, after which the student receives automatic monthly charges unless the subscription is canceled.

- If the student clicks "access course materials", he or she can view the videos and take the quizzes for free but without access to coaching support, a verified certificate, or final project submissions for feedback, all of which are available under the paid version.

Udacity experimented with a free trial screener, where if the student clicked "start free trial," he or she was asked how many hours per week they would commit to the course:

- If the student indicated 5+ hours per week, he or she would be taken through the checkout process.

- If the student 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 suggest the student access the course materials for free.

The hypothesis was the free trial screener sets clearer expectations for students upfront, ultimately reducing the number of frustrated students who leave the free trial. If the hypothesis holds true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

Experiment Design

The goal of the experiment is to decrease the number of unprepared or uncommitted students from enrolling without decreasing revenue. In other words, the screener should adequately filter out the students who will enroll but drop out before the free trial period ends, without making the first payment.

Metric Choice

Invariant metrics: number of cookies, number of clicks, click-through-probability

Invariant metrics are metrics that should not change across experiment and control groups. Because these metrics are “invariant,” they cannot be used as evaluation metrics to determine if the change being tested had an effect on the metric.

*Number of cookies***:** The number of unique cookies to view the course overview page

- The initial unit of diversion is a cookie, so this metric should be evenly distributed across experiment and control groups, allowing it to be used as invariant metric.

*Number of clicks***:** The number of unique cookies to click the "start free trial" button

- Because the change being tested (free trial screener) occurs only after a user has clicked on the “start free trial” button, the probability of a student clicking on the “start free trial” remains unchanged, and hence, the number of clicks can be used as an invariant metric.

*Click-through-probability***:** The number of unique cookies to click the "start free trial" button divided by number of unique cookies to view the course overview page

- Because neither the probability of a unique cookie clicking on the “start free trial” button or the number of unique cookies visiting the course overview page should be affected by the change, the click-through-probability should be consistent across experiment and control groups and can be used as an invariant metric.

Evaluation metrics: gross conversion, retention, net conversion

Evaluation metrics are expected to change throughout the experiment and indicate different responses to the change being tested between the control and experiment groups.

*Gross conversion***:** 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

- If the screener effectively screens out or discourages students who are unlikely to complete the course, the gross conversion rate is expected to decrease for the experiment group. Users who click the “start free trial” button and are prompted about the weekly time commitment may hesitate to complete checkout and enroll if they are certain they cannot commit at least 5 hours/week to the course.

*Retention***:** The 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 checkout

- If the screener effectively screens out or discourages students who are unlikely to complete the course, the retention rate is expected to increase for the experiment group – the students who completed checkout are more likely to commit to completing the course and remain enrolled past the free trial period and make monthly payments.

*Net conversion***:** 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

- If the screener effectively screens out or discourages students who are unlikely to complete the course, the net conversion rate is expected to increase for the experiment group – the users who still choose to enroll after submitting their weekly time commitment to the screener should be dedicated students who are more likely to continue past the free trial period and complete the course.

The final metric, number of user-ids (number of users who enroll in the free trial), was not selected as an invariant metric or evaluation metric. Because this metric would not be distributed evenly between control and experiment groups, it cannot be used as an invariant metric. Because the number of user-ids is a total count that is not normalized and because it is already incorporated into the gross conversion, retention, and net conversion ratios, it was not selected as an evaluation metric.

The final launch criteria for the screener is that all evaluation metrics show statistically significant results, indicating that the screener is indeed achieving the experiment’s goals. In other words, the results of the experiment must show a significant decrease in gross conversion (students who are not prepared to make the time commitment to complete the course will not enroll) and a significant increase in net conversion (the students who do enroll will remain enrolled past the free trial period).

Measuring Standard Deviation

Given a sample size of 5,000 cookies visiting the course overview page:

analytic estimate of standard deviation, σ = $\sqrt{\frac{p\ (1-p)}{N}}$

	Gross conversion	Retention	Net conversion
probability, p	0.2063	0.5300	0.1093
scaled sample size, N	400	82.5	400
standard deviation, σ	**0.0202**	**0.0549**	**0.0156**

An empirical estimate of the variability is needed if the unit of analysis differs greatly from the unit of diversion. Since the unit of diversion is a cookie and both gross conversion and net conversion have number of unique cookies that click “start free trial” as their units of analysis, the analytic and empirical estimates of variability for these metrics will be similar. In the case of retention, the unit of analysis is number of user-ids that complete checkout, so an empirical estimate of variability is likely needed if this metric is used in the final analysis.

Sizing

Number of Samples vs. Power

The Bonferroni correction is a conservative method for controlling Type I errors and can be useful for catching a false positive (statistically significant metric) when measuring multiple metrics. Since the change being tested in this experiment will not likely be deployed unless all evaluation metrics show a statistically significant change, there is little need for the Bonferroni correction and it will not be applied to the following analysis.

	Gross Conversion	Retention	Net Conversion

Baseline conversion rate	0.2063	0.5300	0.1093
Minimum detectable effect, d_{min}	0.01	0.01	0.0075
Statistical power, 1-β	0.8	0.8	0.8
Significance level, α	0.05	0.05	0.05
Sample size per group	25,835	39,115	27,413
Total sample size (2 groups - control & exp)	51,670	78,230	54,826
Unit of analysis (click or enrollment) / pageview	0.08	0.0165	0.08
Number of pageviews needed	**645,875**	**4,741,212**	**685,325**

Total number of pageviews needed (maximum): 4,741,212

Duration vs. Exposure

Implementing a screener to provide expectations on time commitment for a course is low risk and is not expected to cause significant dips in enrollment or user satisfaction. Therefore, it is reasonable to divert or expose 100% of the traffic to the split test. Based on the number of pageviews needed from the previous calculations and an average of 40,000 pageviews per day, the experiment would need to last 119 days, which is unreasonably long and increases risk of exposure and potential opportunity costs. If the fraction of traffic diverted is 50%, the length of the experiment, driven by the retention rate, is still 59 days. If retention is removed as an evaluation metric, then the number of pageviews needed (now driven by net conversion) drops to 685,325, and the length of experiment reduces down to 17 days, a much more reasonable time frame.

	Gross Conversion	**Retention**	**Net Conversion**
% traffic diverted to experiment	100%	100%	100%
length of experiment (days)	16	119	17

Experiment Analysis

Sanity Checks

Number of cookies:

	Control	**Experiment**
Number of pageviews	345,543	344,660
expected fraction of diversion	0.5	
observed fraction of diversion	**0.5006**	
standard deviation, SE	0.0006	
z-score (95% CI)	1.96	
*margin of error, z*SE*	0.0012	
95% confidence interval	**(0.4988, 0.5012)**	
passes sanity check?	**yes**	

Number of clicks:

	Control	**Experiment**
Number of pageviews	28378	28325
expected fraction of diversion	0.5	
observed fraction of diversion	**0.5005**	
standard deviation, SE	0.0021	
z-score (95% CI)	1.96	
*margin of error, z*SE*	0.0041	
95% confidence interval	**(0.4959, 0.5041)**	
passes sanity check?	**yes**	

Click-through probability:

	Control	Experiment
# clicks / # pageviews	0.08213	0.0822
expected fraction of diversion	0.0821	
observed fraction of diversion	**0.0822**	
standard deviation, SE	0.0005	
z-score (95% CI)	1.96	
margin of error, z*SE	0.0009	
95% confidence interval	**(0.0812, 0.0830)**	
passes sanity check?	**yes**	

All invariant metrics pass sanity checks.

Result Analysis

Effect Size Tests

Gross Conversion:

	Control	**Experiment**
enrollments / clicks	0.2189	0.1983
d_{min}	(-)0.01	
observed difference	**-0.0206**	
variance	9.88658E-06	9.21142E-06
standard deviation, SE	0.0044	
z-score (95% CI)	1.96	
*margin of error, z*SE*	0.0086	
95% confidence interval	**(-0.0291, -0.0120)**	
statistically significant?	**Yes, CI does not contain zero**	
practically significant?	**Yes, CI does not contain d_{min}**	

Net conversion:

	Control	**Experiment**
payments / clicks	0.1176	0.1127
d_{min}	(-)0.0075	
observed difference	**-0.0049**	
variance	5.99903E-06	5.79314E-06
standard deviation, SE	0.0034	
z-score (95% CI)	1.96	
*margin of error, z*SE*	0.0067	
95% confidence interval	**(-0.0116, 0.0019)**	
statistically significant?	**No, CI contains zero**	
practically significant?	**No, CI does contain (-)d_{min}**	

Effect size tests show gross conversion is both statistically and practically significant and net conversion is neither statistically or practically significant.

Sign Tests

	Gross Conversion	**Net Conversion**
# successes	4	10
# trials	23	23
2-tailed p-value	**0.0026**	**0.6776**

statistically significant? *(α = 0.05)*	**Yes**	**No**

Sign tests show that gross conversion is statistically significant and net conversion is not statistically significant at α-level 0.05.

Summary

In this experiment, the control group did not receive a screener prompting them to state a weekly time commitment for the course when they clicked on the “start free trial” button. The experiment group was exposed to the screener when they clicked the “start free trial” button. The null hypothesis is that there would be no difference shown by evaluation metrics between control and experiment groups. The unit of diversion was a cookie. The number of cookies and the number of clicks were selected as the invariant metrics. The gross conversion and the net conversion were selected as the evaluation metrics.

The Bonferroni correction method was not used since all evaluation metrics needed to show statistical significance prior to deployment of the change. If the acceptance criteria required only one of the evaluation metrics to show statistical significance and there was a need to control Type I errors more tightly, then the Bonferroni correction method would have been applied.

The final analysis showed that gross conversion was statistically significant at a 95% confidence interval, and therefore rejection of the null hypothesis, but net conversion was not statistically significant. Both effect size tests and sign tests supported these conclusions. It appears the screener successfully diverted some students from enrollment after clicking the “start free trial” button; however, the experiment did not show strong evidence that the screener successfully diverted the students who would not remain past the free trial period, or the students who were taking advantage of the free trial for 14 days and canceling their subscription before making the first payment.

Recommendation

Because net conversion was not statistically significant and failed to reject the null hypothesis, I would recommend not launching the screener. The screener appeared to decrease overall enrollment, apparent in the statistically significant gross conversion rate, but was unable to filter out students who would not fully commit to the course past the free trial period and follow through to course completion. The net conversion confidence interval contained the negative of the practical significance boundary and showed a decrease in the experiment group for number of users who remained past the free trial period and made the first payment relative to the number of cookies who clicked the “start free trial” button. This shows negative business impact because some students in the experiment group who may have gone on to make that first payment (as shown by the control group) are canceling before the free trial period expires.

Rather than launching this screener, I would recommend pursuing further experiments to better filter out the students who will complete the course and utilize the coaching resources provided by Udacity.

Follow-Up Experiment

A future experiment can focus on providing students with even more clarity upfront on what skills and pre-requisites are needed for the course, reducing the risk that students get frustrated after enrolling and give up during the free trial phase.

Setup: Revise the screener to provide a mini-challenge to students when they click the “start free trial” button. The mini-challenge will test basics skills (i.e. programming or math) to give students a better sense of what skills and pre-requisites they will need to complete the course smoothly. If students do not perform well in the mini-challenge, Udacity can refer them to the pre-requisite courses they need and suggest they access course materials for free. Students who do pass the mini-challenge can complete checkout and enroll in the free trial.

Null Hypothesis: no difference in evaluation metrics between the control group (Udacity as-is) and experiment group (mini-challenge screener)

Unit of Diversion: cookie, tracked by user-id after enrollment

Invariant Metrics: number of cookies, number of clicks on “start free trial button”

Evaluation Metrics: gross conversion (# enrollments / # clicks), net conversion (# enrollments past 14-days / # clicks)