Experiment Design

Metric Choice

Invariant Metrics:

Number of cookies: I included number of cookies because this occurs before the screener and is the same diversion level as the evaluation metrics I chose. (d_{min}=3000)

Number of clicks: I included number of clicks because this occurs before the screener and is the same diversion level as the evaluation metrics I chose.(d_{min}=240)

Click-through-probability: I included this metric because it will help reinforce the result of the sanity check for number of clicks.(d_{min}=0.01)

Evaluation Metrics:

Gross conversion: I included gross conversion because this is the metric that the hypothesis predicts will decrease. The unit of diversion is “cookie” which is the intended unit of diversion. If the experiment is to be launched this value must decrease. (d_{min}= 0.01)

Retention: I included retention because this would be an interesting metric to compare with gross conversion, as it will signify the retention of users already enrolled when they saw the screener. If the experiment is to be launched, this metric must not decrease. (d_{min}=0.01)

NOTE: Did not include due to larger number of pageviews needed.

Net conversion: I included net conversion because this metric shows the proportion of students who continue with payment given enrolment. This is important to track because if this decreases, the screener is discouraging students who would have continued with the course. If the experiment is to be launched this value must not decrease. (d_{min}= 0.0075)

Not Used:

Number of user-ids: This could be an evaluation metric, as it would track the amount of students who move past the free trial. However, I did not use number of user IDs because the metric is not normalized, and effects will be difficult to interpret.

Measuring Standard Deviation

Sample cookies visiting course overview page:	5000

Standard Deviation of Evaluation Metrics

Evaluation Metric	Probability	Standard Dev.
Gross Conversion: enrolling, given click:	0.20625	0.0202
Retention: payment, given enroll:	0.53	0.0549
Net Conversion: payment, given click:	0.1093125	0.0156

For Gross Conversion and Net Conversion I believe the analytic estimate will be comparable to the empirical variability due to the fact that unit of diversion is the same as the unit of analysis. For Retention I believe empirical will be the better estimate to use due to the unit of diversion (User-id) being less specific than the unit of analysis, this is usually the case due to the unit of diversion being correlated to the unit of analysis.

Sizing

Number of Samples vs. Power

I choose not to use the Bonferroni correction.

Constants:

alpha	0.05
beta	0.2

Page Views Chart

Evaluation Metric	baseline conversion rate	Dmin	Page Views / variation	Page Views	Total exp + control
Gross Conversion:	0.20625	0.01	25835	322937.5	645875
Retention:	0.53	0.01	39115	2370606.061	4741212.121
Net Conversion:	0.1093125	0.0075	27411	342637.5	685275

The pageviews required are: 645,875 for Gross Conversion, 4,741,213 for Retention and 685,275 for Net Conversion. Since we won’t be using Retention as a metric, we require 685,275 page views.

Duration vs. Exposure

I removed retention as a metric due to the high length of experiment time required. For retention to be a viable metric, i could have increased dmin alpha or beta, but since those are prescribed i chose not to. The unit of diversion is important due to the definition of the metric, if it were changed to cookies vs. unique users it would just be net conversion. The experiment could be targeted to specific traffic, however that would require further exploration of the data.

I chose to divert 80% of all traffic to the experiment, this means it would take 22 days to complete the experiment. This experiment wouldn't be very risky for Udacity, as this will only affect users who state they are able to contribute under 5 hours of their time per week to the coursework. This may either motivate these individuals to contribute more time to the course, or keep them from enrolling. There is also the possibility that individuals who are unable to commit that much time are much less likely to enroll anyway. Because the experiment isn't very risky, I feel confident to divert this much traffic to the experiment.

Experiment Analysis

Sanity Checks

Counted Invariant Observed 95% confidence interval check

invariant metric (count)	total control	total exp	SD	Z-score	m	low bound	upper bound	observed
Unique cookies to view page per day:	345543	344660	0.0006	1.96	0.0012	0.4988	0.5012	0.5006
Unique cookies to click "Start free trial" per day:	28378	28325	0.0021	1.96	0.0041	0.4959	0.5041	0.5005

From the chart, we see that cookie page views and cookies clicking "start free trial" are within the confidence interval and pass our sanity check.

To apply the pooled standard error equation for the proportional invariant N*p has to be greater than 5 and N(1-p) has to be greater than 5. This is true, so we can continue.

Proportional Invariant Observed 95% confidence interval check

invariant metric (probability)	p- pool	Se pool	d	Z-score	z-score * SE
Click-through-probability on "Start free trial":	0.0822	0.00066	0.0001	1.96	0.0013

From the chart, we see that click through probability on "start free trial" is within the confidence interval (m > d ; 0.0013 > 0.0001) and passes our sanity check.

Result Analysis

Effect Size Tests

Rate Calculations for Evaluation Metrics

	Clicks	Enrollments	Payments	Gross Conversion	Net Conversion
Control	17293	3785	2033	0.2189	0.1176
Experiment	17260	3423	1945	0.1983	0.1127
Total	34553	7208	3978	0.2086	0.1151

95% Confidence Intervals for Evaluation Metrics

	Gross Conversion	Net Conversion
Margin of Error	0.0086	0.0067
D	-0.0206	-0.0049
low bound	-0.0291	-0.0116
upper bound	-0.0120	0.0086
Dmin	-0.01	-0.0075

The 95% confidence interval for the difference between experiment and control groups for Gross Conversion is [-0.0291,-0.0120]. Since the difference interval is greater in magnitude than the practical significance boundary of -0.01, and does not include 0, the difference for Gross Conversion is practically and statistically significant.

The 95% confidence interval for the difference between experiment and control groups for Net Conversion is [-0.0116,0.0086]. Since the difference interval includes the practical significance boundary of -0.0075 and 0, the difference for Net Conversion is not practically and statistically significant.

Sign Tests

Evaluation Metric	Net Conversion	Gross Conversion
Sign test “successes”	4	10
P value	0.0026	0.6776
Statistically Significant	Yes	No

Out of 23 days over which the experiment occurred, assuming a 0.5 probability, the experiment result was greater than the control 4 times for Net conversion and 10 times for

Gross Conversion. This gives P values of 0.0026 for Net Conversion and 0.6776 for Gross Conversion.

I did not use the Bonferroni correction. I would suggest that our evaluation metrics are correlated which would make the Bonferroni correction too conservative. Since the correction controls for false positives, the correction is unnecessary as I will require both metrics to pass to launch. The effect size tests and sign tests are in agreement for both Net Conversion and Gross Conversion.

Recommendation

Note: Since we're comparing day by day, and since "Payment" can occur up to 14 days after "Enrolment" I have assumed the data is a selection from a longer run of the experiment, where students who made payments on Thursday, October 16th had still been exposed to the screener, this would mean the experiment had been active since at least October 2nd.

The results show the hypothesis was correct. The decrease in Gross Conversion was statistically and practically significant, showing the Screener did dissuade some students from continuing with the course when they were potentially ill prepared or wouldn't have enough time to complete the course comfortably. There was no statistically or practically significant decrease in Net Conversion, which also means the experiment didn't show a marked decrease in students sticking with the course. We can interpret this as students that wouldn't have stuck with the course may have been screened by the screener: the intended result. The negative practical significance boundary for Net Conversion is within the confidence interval, meaning that there is a possible practical decrease in the ratio of students paying for courses. Because this is a business that requires students to remain in the course and Net Conversion could potentially decrease, I would recommend to not launch this experiment.

Follow-Up Experiment

Now that we know the screener is effective at screening less committed or unprepared students, let's investigate the effect of adding a message at the end of each lesson, suggesting the availability of 1 on 1 instructor time, and referring the student to the forums and peers for help.
The evaluation metric is probability of course cancellation per user-id logins of the students enrolled in the course, this will give us the "dropout ratio". The unit of diversion is user-ID. This is the only real way to track students who are already enrolled in and paying for courses.

The hypothesis is that the pop up will remind frustrated students that resources exist to aid them, thus signalling frustrated students to be more engaged and ask for help. Ideally this will keep frustrated students in the course. A practically and statistically significant decrease in the “dropout ratio” will show the hypothesis to be true. If this hypothesis held true, Udacity could improve the overall student experience and prevent students from dropping courses.

Resources

http://graphpad.com/quickcalcs/binomial1.cfm
http://www.evanmiller.org/ab-testing/sample-size.html