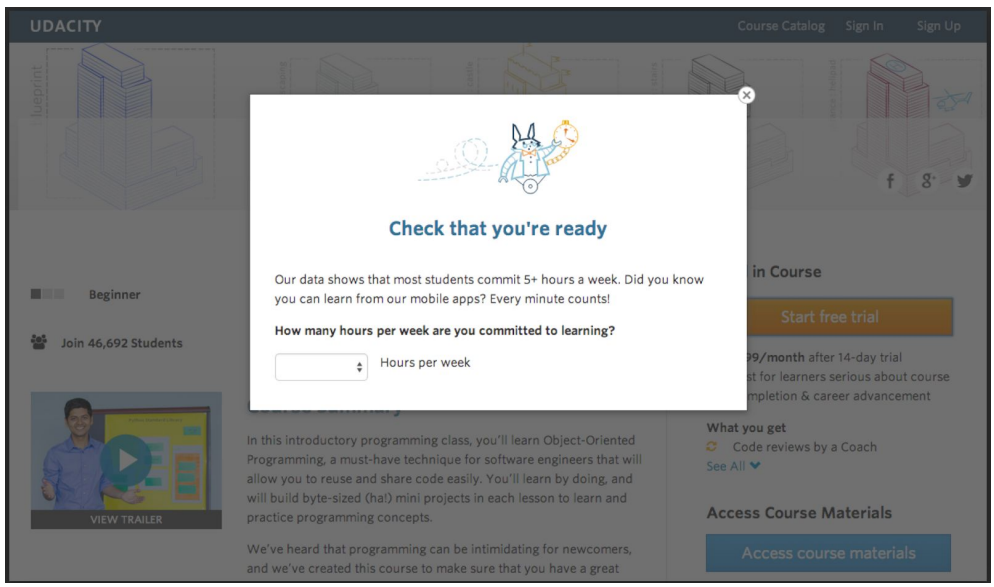
Design an A/B Testing
Nopthakorn Kutawan

Experiment: Udacity Free Trial Screener
Experiment Overview

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

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

Invariant metrics I choose Number of cookies, Number of clicks and Click-through-probability.
Evaluation metrics I choose Gross conversion and Net conversion.

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

(1) Number of cookies:
Invariant metric = Yes, this metric not change with the experiment we used as initial unit of diversion to make sure we had split equally between control and experiment.
Evaluation metric = NO, because it's should not change between control and experiment test.

(2) Number of user-ids:
Invariant metric = No, because the number of enrolled users may depend on the pageview and we may see different between control and experiment.
Evaluation metric = Yes, because we would expect a difference between control and experiment.

(3) Number of clicks:
Invariant metric = Yes, to make sure control and experiment have the same number of clicks, otherwise they have different results because of they have different number of clicks rather than my new feature in experiment.
Evaluation metric = No, because it's should not change between control and experiment.

(4) Click-through-probability:
Invariant metric = Yes, for make sure the population of students are same in control and experiment. Then we can see whether the difference between the results are from new features that we applied.
Evaluation metric = No, because it's should not change between control and experiment.

(5) Gross conversion:
Invariant metric = No, because the experiment had change this metric during experiment.
Evaluation metric = Yes, since same number of clicks for control and experiment, if gross conversion are significantly different, that means new feature make an effect.

(6) Retention:
Invariant metric = No, same as in case of gross conversion
Evaluation metric = No, because this did not influence on experiment.

(7) Net conversion:
Invariant metric = No, because the experiment had change this metric during experiment.
Evaluation metric = Yes, base on the same number of clicks on control and experiment, the new feature will affect the yield of net conversion.

Expectation:
If the hypothesis is correct, we would expect to see significant changes in the evaluation metrics.
To launch the experiment, I expect gross conversion in the experiment is significantly lower than control because those students who likely to drop during the 14-day free trial would be filtered by the new feature. And net conversion would be unchanged because a number of students who continue to pass the free trial and eventually complete the course would not affect.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

In the experiment, we predict that we will need approximately 5,000 cookies per day, from baseline values the estimate of standard deviation (SD) of evaluation metrics = sqrt(p*(1-p)/N).

Gross conversion : SD= sqrt(0.20625*(1-0.20625)/400) = 0.0202
Net conversion : SD = sqrt(0.1093125*(1-0.1093125)/400) = 0.0156

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

Both evaluation metrics (gross conversion, net conversion) the unit of analysis and unit of diversion are same as a number of unique cookies, then I would like to assumption that the analytic estimate should be comparable to the empirical variability.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

Number of samples calculate by web calculator tool (http://www.evanmiller.org/ab-testing/sample-size.html)
The calculation base on alpha = 0.05 and beta = 0.2

	gross conversion	net conversion
Beta	0.2	0.2
Alpha	0.05	0.05
Baseline conversion	0.20625	0.1093125
Dmin	0.01	0.075
Page views from web calculator	25835	27413
Page views divided by CTP	322937.5	342662.5
Total page views	645875	685325

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

Duration

Number of pageviews required (max of all) = 685325
Divert traffic = 100%
Length of experiment = 18 days (685325/40000)

Exposure

The end of an experiment just asks an additional self-assessment question about the time commitment, these question not harms users, also no sensitive information had collected from users. So then, exposure is a 100% safe. Therefore, we could divert the entire traffic to this experiment.

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

The sanity for invariant metric shows the process and values for this

	CIL	CIU	Observed	Passes
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through-probability	0.0812	0.0830	0.0822	Yes

The observed value is within the 95% of CI, all of invariant metric passes the sanity check.

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)

	CIL	CIU	Statistical significance	Practical Significance
Gross conversion	-0.0291	-0.0120	Yes	Yes
Net conversion	-0.0116	0.0019	No	No

The gross conversion is statistical significant and practical significant (zero not CI , but is negative)
The net conversion is not statistical significant and not practical significant (zero include the CI)

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

To calculate the results for the Sign Test I used an online calculator for "Sign and binomial test" (http://graphpad.com/quickcalcs/binomial1.cfm). The results are as follows

	#Success	#Experiments	#Probability	p-value	significance
Gross conversion	4	23	0.5	0.0026	Yes
Net conversion	10	23	0.5	0.6776	No

For gross conversion the 2-tailed p-value 0.0026 is statistically significant ($p<0.05$).
Fo net conversion there was no significant difference ($p>0.05$).

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

I did not use Bonferroni correction in the analysis because our decision is based on Gross Conversion and Net Conversion metrics. And the hypothesis requires two effects be considered in order to launch the experiment rather than the decision in one metric. Therefore, Bonferroni correction will not be appropriate for this experiment.

The results of both tests (Effect Size and Sign Tests) agree with each other, that gross conversion is significant, while net conversion is not significant.
The gross conversion rate was lower in the experimental and thus the new feature was effective to filtered the number of students that enrolled from initial click. This supports our hypothesis.
The net conversion, was decreased but not significant, indicating that the filtered negative effect on the number of students who would be complete the 14-day free trial. Look seen it deterred some students from enrolling that would have otherwise completed the free trial. This is not our intended effect and does not support the hypothesis.

Recommendation

Make a recommendation and briefly describe your reasoning.

Base on hypothesis of this experiment I would conclude that new feature will decreases gross conversion rate by practically significant way. For net conversion rate decreases too but not by a statistically significant.
It seems a new feature was effective in reducing the number of students to continue from click to enrollment but was unable to retain the numbers of students to continue past the free trial unaffected. In fact based upon the confidence interval for the net conversion, it is possible that the time assessment as 5 hours per week might be increased the number of students to leave free trial.

Based on the above observations, I would recommend to not launch this experiment.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

A goal of the experiment is to reduce a frustrated If the student clicks "start free trial".
Based on this goal, I would like to propose the following experiment.
If a student who clicks “Start free trial”, they should perform answer basic questionnaire that help to evaluate the student before they take the course. If they answer some those questions, that mean they interesting in the course so that they should proceed checkout as usual. Otherwise If they don’t perform some questions, a message should appear to indicating that course require basic knowledge for complete the course, by this way I expected it could reduce number of students who left free trial early.

My hypothesis is if students who have some knowledge for the course they may have chance to finish the course, and then a number of students who left the free trial should reduce.

The measuring metrics is a cookie.
Invariant metrics should be Number of cookies, Number of clicks and Click-through-probability for invariant metrics to make equally for control and experiment.
Evaluation metrics should be Gross conversion and Net conversion that could see the effect whether the change influences in experiment.

References

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112991/
http://20bits.com/article/statistical-analysis-and-ab-testing
https://rpubs.com/superseer/abtesting
http://www.evanmiller.org/ab-testing/sample-size.html
http://graphpad.com/quickcalcs/binomial1.cfm