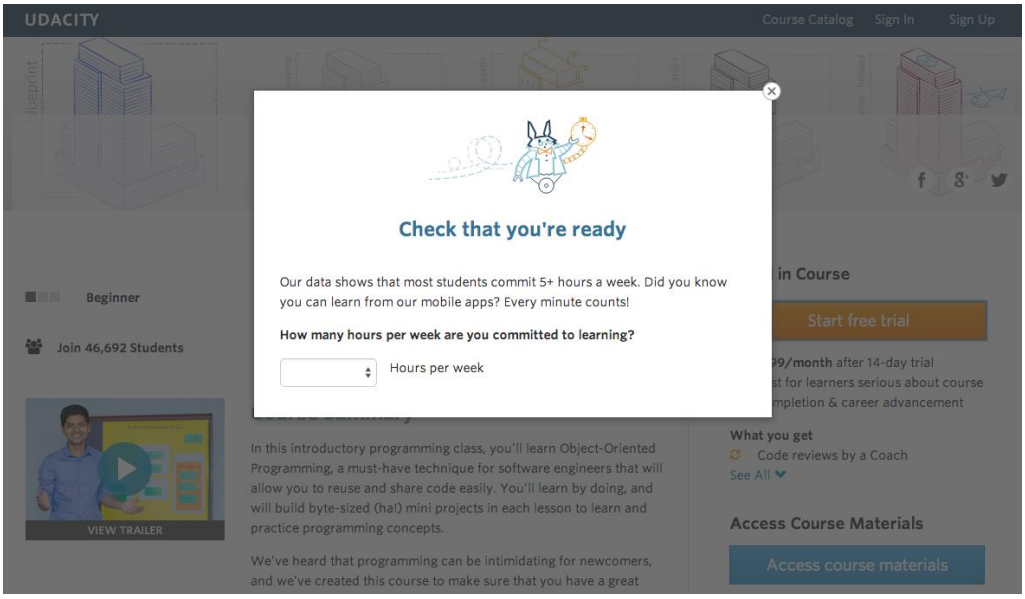
P7: A/B Testing

Experiment Summary

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.

Metric Choice

Invariant metrics should NOT change across experimental or control groups. The following metrics are considered invariants because the intervention occurs AFTER the point where the metrics are measured (therefore the experiment has no direct impact on any of them).

On the other hand, evaluation metrics ARE expected to change over the experiment with differences observed between the experimental and control groups. If the hypothesis is true, the number of user-ids to enroll would be reduced, eliminating the unprepared students while simultaneously keeping payments & checkouts the same. Thus in order to launch the experiment, we want to decrease the enrollment of unprepared students (gross conversion should decrease) while synchronously NOT reducing the number of students who complete the free trial and make a payment (net conversion should either go up or stay the same).

- Number of Cookies: *(Invariant Metric)*

The number of users who visit Udacity's homepage is not contingent upon modifying the Free Trial page

- Number of User Id's *(Evaluation Metric)*

Not very useful as an invariant metric because the number of users who enroll in the free trial is dependent on the experiment. Despite not being the best evaluation metric (not normalized), the number of user id's CAN still be used because it would track the first part of the hypothesis (whether we reduce the number of students who continue past the free trial).

- Number of Clicks *(Invariant Metric)*

Users who have clicked have not yet seen the "5+ hours" commit message from the Start Free Trial page.

- Click-Through-Probability *(Invariant Metric)*

Similar to the previous answer, users who have clicked have not yet seen the "5+ hours" commit message, thus deeming this metric as Invariant.

- Gross Conversion *(Evaluation Metric)*

The number of users who decide to start the free trial are expected to depend on how the start free trial page is rendered (whether a "5+ hours" commit message is suggested), thus this is one question we would like to understand at the end of the A/B test.

- Retention *(Evaluation Metric)*

Not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. It is however a rather good evaluation metric (if the size isn't too big) because it is directly dependent on the effect of the experiment.

- Net Conversion *(Evaluation Metric)*

The net conversion is the product of the previous two evaluation metrics (gross conversion & retention). This can also be considered as a more general goal of the A/B test –whether rendering a "5+ hours" commit message suggestion helps increase the ratio of users who make payments over those who see the start free trial page.

Variability

For Bernoulli distribution with probability p and population N, the analytical deviation is … SD = SQRT(P X (1 – P) / N)

Thus to assess whether analytical estimates of the standard deviation are accurate (whether it matches with the empirical standard deviation), we consider whether or not the unit of analysis and unit of diversion match up). I will use Gross Conversion and Net Conversion metrics for evaluation and try to investigate "Would this test reduce the number of frustrated students who left the free trial because they didn't have enough time?

- Gross Conversion

SQRT (P X (1 – P) / N) = SQRT (0.20625 X (1 – 0.20625) / (5000 X .08)

Standard Deviation = 0.0202

Denominator = Cookie [which is also Udacity's Unit of Diversion]; thus variance CAN be analytically estimated.

If any of the denominators were something different than a Cookie, THEN we'd have to empirically estimate the variance.

- Net Conversion

SQRT (P X (1 – P) / N) = SQRT (0.1093125 X (1 – 0.1093125) / (5000 X 0.08)

Standard Deviation = 0.0156

Denominator = Cookie [which is also Udacity's Unit of Diversion]; thus variance CAN be analytically estimated.

If any of the denominators were something different than a Cookie, THEN we'd have to empirically estimate the variance.

Sizing

I decided NOT to use the Bonferroni correction in my analysis phase because it is too conservative to pick up on metrics that are strongly correlated with each other. (Alpha = 0.05 & Beta = 0.2)

- Gross Conversion

Baseline Conversion Rate = 0.20625 & d_min = 0.01 & Required # Samples = 25835

(25835 / 0.08) X 2 = 645875 Page Views

- Retention

Baseline Conversion Rate = 0.53 & d_min = 0.01 & Required # Samples = 39115

(39115 / 0.08 / 0.20625) X 2 = 4741212 Page Views

- Net Conversion

Baseline Conversion Rate = 0.1093125 & d_min = 0.0075 & Required # Samples = 27413

(27413 / 0.08) X 2 = 685325 Page Views

It appears required page views for our Retention metric is too large (4.7 million) & would require 117 days of full traffic, deeming it impractical. Gross Conversion & Net Conversion however are both sensible for our experiment. Because the required number of page views is higher for our Net Conversion metric (685325 page views), we will choose this figure to work with.

Duration & Exposure

With 685325 page views required, we direct 50% of our traffic to the experiment.

685325 / (40000 X 0.5) = 35 Days to Run the Experiment

I would divert 70% of the total traffic to the experiment, bringing the duration of the experiment to 25 days (which is reasonable for our needs). I do not believe this experiment is particularly risky because it does not affect paying customers and has a low chance of having any bugs (Udacity clients should have little concern of having their limited sensitive information compromised.)

Sanity Checks

- Number of Cookies

Total Control Group = 345543

Total Experiment Group = 344660

Standard Deviation = SQRT (0.5 X 0.5 / (345543 + 344660)) = 0.0006018

Margin of Error = 1.96 X 0.0006018 = 0.0011796

Lower Bound = 0.5 – 0.0011797 = 0.4988

Upper Bound = 0.5 + 0.0011797 = 0.5012

Observed = 345543 / (345543 + 344660) = 0.5006

0.4988 < 0.5006 < 0.5012 = Observed Value Within Bounds (Invariant Metric Passes)

- Number of Clicks on “Start Free Trial”

Total Control Group = 28378

Total Experiment Group = 28325

Standard Deviation = SQRT (0.5 X 0.5 / (28378 + 28325)) = 0.0021

Margin of Error = 1.96 X 0.0021 = 0.0041

Lower Bound = 0.5 – 0.0041 = 0.4959

Upper Bound = 0.5 + 0.0041 = 0.5041

Observed = 28378 / (28378 + 28325) = 0.5005

0.4959 < 0.5005 < 0.5041 = Observed Value Within Bounds (Invariant Metric Passes)

- Click-Through-Probability on “Start Free Trial”

Control = 0.0821258

Standard Deviation = SQRT (0.0821258 X (1 – 0.0821258) / 344660) = 0.000468

Margin of Error = 1.96 X 0.000468 = 0.00092

Lower Bound = 0.0821258 – 0.00092 = 0.0812

Upper Bound = 0.0821258 + 0.00092 = 0.0830

Observed = 0.0821824

0.08212 < 0.0822 < 0.0830 = Observed Value Within Bounds (Invariant Metric Passes)

Effect Size Tests

- Gross Conversion

n_cnt = Clicks Controlled = 17293

x_cnt = Enroll Controlled = 3785

n_exp = Clicks Experiment = 17260

x_exp = Enroll Experiment = 3423

p_pooled = (x_cnt + x_exp) / (n_cnt + n_exp) = 0.2086

se_pooled – SQRT(p_pooled X (1 – p_pooled) X (1 / n_cnt + 1/n_exp)) = 0.00437

d = x_exp / n_exp – x_cnt / n_cnt = -0.02055

lower bound = d – se_pooled = -0.0291

upper bound = d + se_pooled = -0.0120

0 is NOT in the Interval; Metric is Statistically Significant & Practically Significant

- Net Conversion

n_cnt = Clicks Controlled = 17293

x_cnt = Enroll Controlled = 2033

n_exp = Clicks Experiment = 17260

x_exp = Enroll Experiment = 1945

p_pooled = (x_cnt + x_exp) / (n_cnt + n_exp) = 0.1151

se_pooled – SQRT(p_pooled X (1 – p_pooled) X (1 / n_cnt + 1/n_exp)) = 0.00343

d = x_exp / n_exp – x_cnt / n_cnt = -0.0048

lower bound = d – se_pooled = -0.0116

upper bound = d + se_pooled = 0.0019

0 IS in the Interval; Metric is Statistically Insignificant & Practically Insignificant

Sign Tests

- Gross Conversion

Days in Experiment Group We See Improvement = 4/23

Probability for Sign Test = 0.5

P-Value = 0.0026; Change IS Statistically Significant (Less than 0.05 Alpha)

- Net Conversion

Days in Experiment Group We See Improvement = 10/23

Probability for Sign Test = 0.5

P-Value = 0.6776; Change is NOT Statistically Significant (Greater than 0.05 Alpha)

Summary

The Bonferroni correction was NOT used in my analysis. This correction is a method to limit & control the risk of Type 1 errors in multiple independent metric comparisons. If one metric was used to base our decisions then this method would have been appropriate; however, both gross conversion AND net conversion were considered. The sign & effect size tests both showed gross conversion to be statistically significant, while proving net conversion not to be.

Recommendation

Examination of the screener provided by gross conversion indicates that the experiment was successful at reducing the number of unprepared students who enrolled in the free trial.

Net conversion results however are both statistically and practically insignificant and the confidence interval includes NEGATIVE numbers of the practical significance boundary, meaning it's possible this number went down by an amount that would hurt the business (decrease revenue). If we consider the initial hypothesis, it does NOT increase the number of paid users in the end. Given these results, we choose NOT to launch. The risk is not worth the reward.

Follow-Up Experiment

I propose extending the length of the free trial from 2 weeks to 1 month (30 days). This will grant new users an additional two weeks of usage before they are forced to decide to enroll or not (and consequently get charged the standard $200 / month fixed rate). The hypothesis is that by providing new users a little more time, they will have had made enough progress in their studies to want to enroll as a user. Just speaking anecdotally, it took me personally about two weeks just to get enough discipline and courage to start the first optional project. I believe this extends to other users as well.

The units of diversion are user-ids because the experiment is held AFTER students enroll in the free trial.

The invariant metric are the number of user-ids. Creating an account (and getting a unique user-id) is before the enrollment process so we don't expect this to change between control and experiment groups.

The evaluation metric will be retention, which is number of users ids to remain enrolled past the 30 day mark (and thus make at least one payment) divided by the number of user-ids to complete checkout. We want to test whether increasing the length of the free trial from 2 weeks to 4 weeks (30 days) will increase the chances of them enrolling as a user (and thus making their first payment).