

Final Project

March 4, 2016

1 EXPERIMENT DESIGN

"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.

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."

1.1 METRIC CHOICE

- Number of cookies: That is, number of unique cookies to view the course overview page. (dmin=3000)
- Number of user-ids: That is, number of users who enroll in the free trial. (dmin=50)
- Number of clicks: That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (dmin=240)
- Click-through-probability: That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. (dmin=0.01)
- 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. (dmin= 0.01)
- 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 check-out. (dmin=0.01)
- 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. (dmin= 0.0075)

1.1.1 INVARIANT METRICS

There are several invariant metrics that could be used over the course of the experiment. Each of these should not change between the experimental and control groups, and will provide a way to double-check the integrity of our design. Those chosen for the experiment:

- Number of cookies (approximation of unique pageviews)
- Number of clicks
- Click-through-probability

Being as the screener pops-up after clicking on the 'start free trial' button, the number of pageviews, clicks, and click-through-probability should not be affected. Anything after the screener; number of user-id's, gross conversion, retention, and net conversion could be affected.

1.1.2 EVALUATION METRICS

We expect that our evaluation metrics will change over the course of the experiment. By comparing differences between control and experimental groups, we can measure the effect of the screener.

- Gross conversion rate (could measure whether or not the screener had an effect on enrollment)
- Retention rate (could measure whether or not the screener had an effect on the 14-day dropout rate)
- Net conversion rate (could measure whether or not the screener had any effect on the 14-day completion rate, although not able to tell us where in this process)

User-id is excluded as an evaluation metric since it is a count, and gross conversion offers a better way to track the effect of the screener on enrollment.

1.2 MEASURING STANDARD DEVIATION

Before conducting the experiment, data was collected to get daily values for cookies, enrollments, click through probability, gross conversion, retention, and net conversion on Udacity's website. The data collected is referred to as the baseline.

In the experiment, we predict that we will need approximately 5,000 cookies per day in each group. From this, a rough estimate of the expected standard deviation for each evaluation metric can be calculated. First, to get an approximation of the number of clicks and enrollments for this daily sample of 5,000 cookies, we scale by the fraction of pageviews in the sample over the pageviews in the baseline $\frac{5000}{40000} = 0.125$. Therefore, we predict 400 clicks and 82.5 enrollments per day in the sample.

The number of clicks and enrollments follows a binomial distribution, and by the central limit theorem, the distribution of the rates (gross conversion, retention, and net conversion) is gaussian. The standard deviation of these normally distributed rates is $\sigma = \sqrt{\frac{p(1-p)}{n}}$. The rates for the evaluation metrics are:

$$p_{gc} = 0.20625 \quad p_r = 0.53 \quad p_{nc} = 0.1093125$$

Calculation of the standard deviations yields:

$$\sigma_{gc} = 0.0202 \quad \sigma_r = 0.0549 \quad \sigma_{nc} = 0.0156$$

Post experiment, the actual number of cookies used per day is higher than our estimated value. We are at nearly 10,000 cookies per day per group rather than 5,000. Doubling the

expected sample size for clicks and enrolls, we can make a revised analytic estimate for standard deviation. The values that we expect to see are:

$$\sigma_{gc} = 0.0143 \quad \sigma_r = 0.0388 \quad \sigma_{nc} = 0.0110$$

1.3 SIZING

1.3.1 NUMBER OF SAMPLES VS. POWER

To know the exact number of pageviews required for our experiment, we calculate the sample size that we will need for each evaluation metric. Let us assume we allow a type I error rate of $\alpha = 0.05$, and type II error $\beta = 0.20$ for each metric. The minimum detectable effect for each evaluation metric has been prespecified (as a business decision):

$$d_{min}^{gc} = 0.01 \quad d_{min}^r = 0.0075 \quad d_{min}^{nc} = 0.01$$

Using the rates from the baseline sample along with this α and β , a sample size calculator will give us the required number of samples for each evaluation metric. Each metric has it's own unit of size (clicks or enrolls), so once we arrive at the required sample size, we need to scale from the given unit to pageviews by the ratio seen in the baseline multiplying by 2 for both the control and experiment groups:

$$\text{ratio of pageviews to clicks} = 0.08$$

$$\text{ratio of pageviews to enrolls} = 0.0165$$

$$n_{gc} = 645875 \quad n_r = 4741213 \quad n_{nc} = 685275$$

The largest sample size is our limiting factor (retention rate), so we require a total of 4,741,212 pageviews to conduct the experiment.

1.3.2 DURATION VS. EXPOSURE

Given the required pageviews for our experiment, we can specify an exposure and calculate duration of the experiment. Dividing total pageviews by the number of pageviews per day in the baseline (40,000), gives us a duration of 119 days were Udacity to divert it's entire traffic. This is too long an experiment. In order to cut down on duration, we can exclude retention rate as an evaluation metric and consider the next limiting metric, conversion rate. With 685,275 necessary pageviews, we can specify exposure at 50% of all traffic for the experiment. Diverting half the traffic per day (20,000), it would then take 35 days to run the experiment which is a more reasonable duration. Because the screener is a mild reminder about study time, it's effect on overall enrollment should be small, and a 50% diversion should be safe.

To be sure of this, consider a worse case scenario in which the screener lowers retention rate by 1 standard deviation (0.0388). With an expected 165 enrollments per day, half of which are diverted to the experiment and from this a half assigned to the experimental group, 41.25 people would be affected by the screener. Under normal circumstances, 53% of these would go on to make payment for a total of approximately 22 people. Assuming the worst, the rate would drop to 49.12% and only about 21 people would make payment per day. This amounts to the loss of 1 person per day which confirms that a 50% should be safe.

2 EXPERIMENT ANALYSIS

2.1 SANITY CHECKS

Having conducted the experiment, we can double-check our invariant metrics to see if our underlying assumptions are being met. We expect that cookies and clicks are divided evenly between the control and experimental groups. Using an expected rate of diversion of 0.5, we can calculate the standard deviation (using the a normal approximation to the rate as before), and construct a 95% confidence interval around our expected value. Comparing the observed rate, we can check if these two invariant metrics are reliable.

$$p = 0.5 \quad \alpha = 0.05 \quad Z = 1.96$$

$$\sigma_{cookies} = \sqrt{\frac{p(1-p)}{345543+344660}} = 0.00601841$$

$$\text{Margin of Error} = 1.96 * 0.00601841 = 0.001179608$$

$$\text{Confidence Interval} = [0.4988, 0.5012]$$

$$\text{Observed rate} = 0.500639666$$

$$\sigma_{clicks} = \sqrt{\frac{p(1-p)}{28378+28325}} = 0.002099747$$

$$\text{Margin of Error} = 1.96 * 0.002099747 = 0.004115504$$

$$\text{Confidence Interval} = [0.4959, 0.5041]$$

$$\text{Observed rate} = 0.500467347$$

Both cookies and clicks pass the sanity check. For click-through-rate (CTR), we should observe more or less the same value across groups. Using the observed rate in the control group, we can construct a confidence interval, but instead compare the observed rate in the experiment group. This will test whether or not the two rates come from the same population which is what we would expect.

$$CTR_c = 28378 / 345543 = 0.082125813$$

$$\sigma = \sqrt{\frac{0.0821(1-0.0821)}{345543}} = 0.000467$$

$$\text{Margin of Error} = 1.96 * 0.000467 = 0.00091532$$

$$\text{Confidence Interval} = [0.0811, 0.0830]$$

$$CTR_x = 28325 / 344660 = 0.0822$$

CTR passes the sanity check, and all of our invariant metrics appear to be well chosen.

2.2 RESULTS ANALYSIS

2.2.1 EFFECT SIZE TESTS

For each evaluation metric, we test for statistical and practical significance (whether or not the size of the effect is relevant from a business standpoint). The minimum detectable effect is the smallest difference that we will accept between experimental and control groups in order to be practically significant. Using the data collected, we calculate the rate in experimental and control groups for each evaluation metric (gross conversion, net conversion), and then define a new variable that is the difference between the rates (experiment - control). Using this newly defined variable, we construct a confidence interval which will then set a range for the expected difference. Because we are using multiple evaluation metrics to test our experimental hypothesis, we use the Bonferroni correction to determine the individual type I error for each evaluation metric:

$$\alpha_{ind} = \frac{\alpha_{total}}{n} = 0.025$$

$$\alpha_{ind/2} = 0.0125 \text{ (two-sided test)}$$

$$Z = 2.24$$

Gross Conversion

$$r_c = 0.2188746892 \quad r_x = 0.1983198146 \quad \hat{d} = -0.020554874$$

$$Var_c = \frac{0.2188746892 * (1 - 0.2188746892)}{17293} = 9.88657605 * 10^{-6}$$

$$Var_x = \frac{0.1983198146 * (1 - 0.1983198146)}{17260} = 9.211417482 * 10^{-6}$$

$$Var_{\hat{d}} = 9.88657605 * 10^{-6} + 9.211417482 * 10^{-6} = 1.909799252 * 10^{-5}$$

$$\sigma_{\hat{d}} = 0.004370125$$

$$ME = 9.7888 * 10^{-3}$$

$$CI = [-0.0303, -0.0108]$$

$$d_{min} = (-)0.01$$

Net Conversion

$$r_c = 0.1175620193 \quad r_x = 0.1126882966 \quad \hat{d} = -0.004873723$$

$$Var_c = \frac{0.1175620193 * (1 - 0.1175620193)}{17293} = 5.999027983 * 10^{-6}$$

$$Var_x = \frac{0.1126882966 * (1 - 0.1126882966)}{17260} = 5.793142782 * 10^{-6}$$

$$Var_{\hat{d}} = 5.999027983 * 10^{-6} + 5.793142782 * 10^{-6} = 1.179217076 * 10^{-5}$$

$$\sigma_{\hat{d}} = 3.43397029 * 10^{-3}$$

$$ME = 7.692099585 * 10^{-3}$$

$$CI = [-0.0126, 0.0028]$$

$$d_{min} = (+)0.01$$

Gross conversion is both statistically and practically significant but net conversion is neither.

2.2.2 SIGN TESTS

To further test each of our evaluation metrics, we can conduct a binomial sign test. Each day of the experiment is evaluated to see if there is a positive or negative difference across groups (experimental - control). We count each positive difference as a success, and each negative difference as a failure. Comparing the resulting p-values for each metric, we can determine significance. Gross conversion rate has 4 of 23 successes for a two-tailed p-value of 0.0026. This is much smaller than our individual type I error (from the Bonferroni correction) of 0.025 indicating statistical significance of gross conversion. Net conversion has 10 of 23 successes and a two-tailed p-value of 0.6776 indicating that net conversion is not statistically significant.

2.2.3 SUMMARY

The Bonferroni correction was used in the analysis phase in order to make a conservative estimate of the individual type I error rates. Any discrepancies between the sign tests and the effect size tests would warrant further study. In this case, the sign tests seem to agree with our effect size tests that gross conversion is significant and net conversion is not.

2.3 RECOMMENDATION

In conclusion, we should not launch the change as although gross conversion was significant, net conversion was not. This implies that the screener deterred people from enrolling in a free trial, however, there was no positive effect on the 14-day trial dropout rate. Essentially, the screener was not effective at keeping out those people that would drop before the 14-day trial, and keeping those in that would proceed past it.

3 FOLLOW-UP EXPERIMENT

A follow-up experiment could be based upon motivation and engagement. The experiment would require a method to approximate the number of hours that each student dedicated to the material per week. If a student committed less than the recommended number of hours in the first week, a message would pop-up upon login before the start of the second week to

motivate the student to commit more time. Those that met the recommended number of hours wouldn't get a message. At this point, my hypothesis is that the message would motivate some students to continue past the 14-day trial and possibly complete the course, but also cause some students to drop-out before that start of the second week. If the first is true, we may be able to motivate students to change their habits. If the second is true, those students that may not complete the program would then leave. In this case, the overall student experience in the forums could improve beyond the second week, and coaching resources could be used on more dedicated students.

Considering this design, we could include an additional evaluation metric, the '7-14 day' retention rate. Both the units of diversion and analysis would be user-id as in the retention rate. The overall retention could allow us to evaluate if the pop-up message had any effect. The '7-14 day' retention rate would evaluate whether or not the pop-up message had an effect to motivate students to continue past the 14-day trial period. This experiment could involve more risk for Udacity than the previous, and the amount of traffic diverted would need to be smaller than 50%. It would also likely take longer to conduct this kind of experiment, recalling that retention rate required significantly more time than gross or net conversion rates.