Gary Ng
DAND *Project 7 – A/B Testing on Udacity*

Experiment Design

Metric Choice

My choice of evaluation and invariant metrics for experimental design are summarized below:

Metric	Choice	Rationale	Expectation
Gross conversion	Evaluation metric	Gross conversion measures the proportion of interested students who clicked on "start free trial" that completed the checkout process and enrolled in the free trial. This is a suitable evaluation metric for the first part of the hypothesis which states, "…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," which effectively postulates that the prompt would successfully discourage some from enrolling the free trial thus lowering the gross conversion rate.	Gross conversion for experiment group to be lower than that for control
Net conversion	Evaluation metric	Net conversion measures the proportion of interested students who clicked on "start free trial", that not only enrolled in the free trial but remained enrolled through the 14 days to become paying members. This is a suitable evaluation metric for the second part of the hypothesis which states, "… without significantly reducing the number of students to continue past the free trial," which effectively postulates that the prompt would not turn away students that will ultimately complete the free trial and become paying members (effectively keeping net conversion the same), even if gross conversion might have suffered as "less committed" students are discouraged by the prompt.	Net conversion for experiment group to be at least equal to that of control
Number of user-ids	Neither (Evaluation metric if have to)	This is the numerator of gross conversion, and thus could be used for an evaluation metric to test the first part of the hypothesis as discussed earlier. However, since this is not normalized, it is not the best metric to use. Instead, we will rely on gross conversion to assess the same effect.	(If used: number of user-ids in experiment group to be lower than that in control)
Number of cookies	Invariant metric	The unit of diversion is cookie, thus by virtue of experimental design, this is randomized and expected to be invariant.	Even split (50%) of cookies between control and experiment

			groups
Number of clicks	Invariant metric	Up to the point when users clicked on “start free trial”, the experiment and control conditions are identical (since users are not shown the “time commitment” prompt until post-click). Thus there is no reason to expect number of clicks to vary and thus makes for a good choice for invariant metric.	Even split (50%) of clicks between control and experiment groups
Click-through-probability	Invariant metric	Click-through-probability is derived from number of clicks and number of cookies. Since these two metrics are not expected to vary, neither is click-through-probability and thus it should be an invariant metric.	No difference in click-through-probability between control and experiment groups
Retention	Neither	Retention measures the percentage of users who enrolled in the free trial that ultimately became paying members (i.e. lasted 14 days). The “time commitment” prompt was precisely designed to increase this percentage, thus it should have been a good candidate for an evaluation metric. However, we computed that doing so would require an impractically high N (months of Udacity web traffic even if 100%-directed), thus we decided to not use this as an evaluation metric. Moreover, one could argue that because we are not concerned with retention rate in a vacuum (e.g. 100% retention would still prove detrimental to the business if the “time commitment” prompt deterred majority of people from enrolling at all), it is acceptable to exclude retention as an evaluation metric. However, it should not be an invariant metric since we do not expect nor hope for it to be constant.	N/A

Measuring Standard Deviation

Given a sample size of 5000 cookies visiting the course overview page, we computed the analytic estimate of the standard deviation for gross conversion to be 0.0202, and that for net conversion to be 0.0156.

As the unit of analysis and unit of diversion are the same (i.e. cookies), we expect the analytic estimates to be comparable with the empirical variability. Having said that, given time and resources, it will always be good practice to corroborate our assumptions with A/A tests (bootstrapping or otherwise).

Sizing

Number of Samples vs. Power

In order to meet the minimum detectable difference (if there was indeed an effect) as prescribed by Udacity's practical significance thresholds for gross conversion (1%) and net conversion (0.75%) rates, we need at least 645,975 and 685,275 cookies respectively. Thus, the number of page views required to power the experiment is the higher of the two: 685,275. Note that Bonferroni correction was not applied in the calculations.

Evaluation Metric	Gross Conversion	Net Conversion
P_hat	0.2063	0.1093
SD	0.0202	0.0156
d_min (practical significance)	0.0100	0.0075
Clicks per variation needed	25,839	27,411
Total cookies needed*	645,975	685,275

** Note: Total cookies needed was derived from multiplying clicks per variation by two (control and experiment) and then dividing by click-through probability of 0.08*

Duration vs. Exposure

Portion of traffic diverted	20%	30%	40%	50%	60%	70%	80%	20%
No. of days needed	86	58	43	35	29	25	22	86

Assuming Udacity's baseline traffic of 40,000 unique visits per day, the table above shows the number of days needed to run the experiment given specific levels of traffic diverted. I would argue that the risk of this experiment is relatively low. Granted, there exists some downside risk that increased friction might prevent some potential paying customers from following through with the enrollment process altogether. However, the financial impact is likely limited, amounting to only the lifetime value of few marginal customers potentially lost during the course of the experiment that might have otherwise enrolled and eventually paid, and Udacity can easily reverse its course of action if the results appear to be unfavorable. As such, Udacity should feel comfortable diverting a significant portion of its traffic to this experiment. I would recommend using 50% of its traffic to close out the experiment in about 5 weeks, which seems like a reasonable amount of time for a non-urgent A/B test. Note that I chose not to divert more than 50% (despite the relative low risk) in order to reserve some traffic in case additional A/B tests are to occur concurrently.

Experiment Analysis

Sanity Checks

The 95% confidence interval for the invariant metrics, along with their actual observed values are computed below. All three invariant metrics are within the confidence intervals and thus pass our sanity check.

Invariant metric	Lower bound of 95% CI	Upper bound of 95% CI	Observed Value	Passes Sanity Check
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of user-ids	0.4959	0.5041	0.5005	Yes
Click-through-probability*	-0.0013	0.0013	0.0001	Yes

** Note: For click-through probability, the 95% confidence interval was built for the difference in proportion between the control and experiment groups. For all other (count) invariant metrics, confidence interval was built around the fraction of events assigned to control group.*

Result Analysis

Effect Size Tests

The effect size estimates along with their 95% confidence intervals are computed below for both evaluation metrics. The gross conversion results were consistent with our expectations, yielding decreases that were both statistically and practically significant. On the other hand, results from net conversion were more concerning. While the decrease in net conversion was neither statistically nor practically significant (i.e. confidence interval covering both 0 and -0.0075), the entire confidence interval was in the negative region, with the upper bound (i.e. worst case scenario) far exceeding Udacity's practical significance of -0.0075. This means that there is a decent likelihood that the prompt might have a detrimental effect on net conversion that is deemed material by the business.

Evaluation metric	Effect size	Lower bound of 95% CI	Upper bound of 95% CI	Practical Sig. Threshold	Statistically Significant	Practically Significant
Gross conversion	-0.0206	-0.0291	-0.0120	-0.0100	Yes	Yes
Net conversion	-0.0049	-0.0116	0.0019	-0.0075	No	No

** Note: Difference (i.e. effect size) is expressed in terms of Experiment group minus Control group*

Sign Tests

We also ran a sign test using the day-by-day data. Of the 23 days for which we could compute the eventual conversion rates, the experiment condition outperformed control in 10 days for net conversion, but only 4 days for gross conversion. This gives rise to a p-value of 0.6776 (not significant) for the former, and 0.0026 (significant) for the latter.

Evaluation metric	# of Days Experiment outperformed Control	Total # of Days	p-value	Statistically Significant
Gross conversion	4	23	0.0026	Yes
Net conversion	10	23	0.6776	No

Summary

The sign tests corroborated our findings in the effect size hypothesis tests. We found that the "time commitment" prompt resulted in a statistically significant drop in gross conversion, but not in net conversion. Note that our calculations did not use a Bonferroni correction given that the results from both our hypotheses were critical in our decision to move forward with the full launch.

Recommendation

Our working hypotheses were that the experimental prompt will dissuade some "less committed" students from enrolling in the free trial that help free up capacity, while keeping the proportion of students who ultimately enroll as paying members relatively constant. Our A/B test results have shown that while the prompt did succeed in reducing the proportion of visitors who enroll in the free test (gross conversion), it has also hurt our net conversion to an extent that the business might deem concerning. There is a high risk that Udacity might suffer a loss in revenue if it went ahead with the change; as such, I would recommend that Udacity not launch the change at this point.

Follow-Up Experiment

Perhaps Udacity can adopt a different strategy to reduce the number of frustrated students who cancel early in the course. Instead of creating an "expectation-setting" message at the onset that might turn away prospective paid users, Udacity should consider experimenting with the user experience post-enrollment to curb churn instead.

One way to do this would be to display career-oriented interstitials (no more than once every three days, lest it becomes annoying) to enrolled students upon login during their 14-day free trials. Each interstitial will feature one Udacity graduate who has successfully transitioned into a new role related to and presumably made possible by his Nanodegree. Since a user may be exposed to more than one interstitial, Udacity should create multiple creative versions and code A/B test logic so that no user is shown the same interstitial twice. The suitable unit of diversion is user id since we are testing some sort of user behavior conditioning post-enrollment, thus each user should either be shown interstitials or not at all.

The hypothesis is that users might be motivated by the success stories of the Udacity graduates featured to persevere in their learning and are thus more likely to remain in the course past the

free trial period. Moreover, it also serves as the “last line of defense” that users would have to get through before making the cancelation (since they have to login first).

The primary evaluation metric will be retention, defined as the number of user-ids that remain enrolled past the 14-day mark (and thus make at least one payment) divided by number of user-ids to complete enrolment checkout. This measures the primary motivation of the proposed experiment, which is to reduce churn and in turn improve retention. However, it should be caveat that we learnt from the past experiment that a large sample might be needed to measure retention to the practical significance level typically mandated by the company. Thus to complete this experiment, Udacity might have to consider lowering that threshold and / or conduct it over a longer period of time than usual. As sanity checks, I would choose number of user-ids assigned (to control vs. experiment groups) as well as unique daily views of the course page by the two groups. The former is the unit of diversion, thus by experimental design ought to be randomized should the systems work correctly. The latter ought not to vary since user ids were distributed randomly, but we should check that we did not have one group that is intrinsically more motivated (i.e. visiting course page more frequently) than the other as that would bias the study. To move forward with the experiment, we would want to see higher retention in the experiment group relative to control, as well as no discernable difference in the two invariant metrics.