Udacity A/B Test Write-Up

Eric Gordon

Experiment Design

Metric Choices

Invariant Metrics:

1. Number of Cookies - The number of unique cookies to view the course overview page.
2. Number of Clicks -The, number of unique cookies to click the "Start free trial" button.
3. Click Through probability- The number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.

To make sure we have an equal split in our control and experimental group we want to check if of our experimental unit of diversion is properly assigning individuals to the two groups. Thus the number of cookies, which is our unit of diversion, will be the first invariant metric in this experiment. We need an additional invariant metric in my opinion though, because the experimental component of this website does not prompt until an individual clicks the “Start Free Trial” button. Thus one another invariant metric we will check is “Number of clicks.” That is, we want to make sure that the experimental and control groups have about the same number users who end up clicking the “Start Free Trial” button. This is vital because about the same number of individuals need to be prompted with the experimental pop-up, which requires a click. Additionally then, we will check the click through probability as our final invariant metric, to make sure the probability of clicking the button for a web page view is about the same for the control and experimental group. Past this point, the metrics we measure will have to be evaluation metrics.

Evaluation Metrics:

1. Net Conversion - The number of user-ids to remain enrolled past the 14-day boundary divided by the number of unique cookies to click the "Start free trial" button.
2. Gross Conversion -The 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.

For this experiment, we want to measure whether or not adding a time commitment warning will decrease the number of students who leave the free trial and do not complete the class, while also not reducing the number of students who make it past the free trial. So we need at least two different evaluation metrics to draw conclusions from this experiment.

The first evaluation metric we will use, is the Gross conversion of students. This metric will be used to make sure that the overall number of students who enroll in the class does not decrease because of our experimental time commitment warning. If this rate significantly decreases from our experimental group, we will have to consider that the experimental prompt is deterring students from even starting the course, a result we do not want.

Our second metric we will use to measure the dropout rate will be the Net conversion. This will help measure our rate of students who drop out of the class during the free trial. If this rate increases significantly, we may be able to conclude that our time commitment warning is helping raise awareness of the time commitment, and we may be able to reject our null hypothesis.

On a quick note, a measurement of pure retention was almost considered as an evaluation metric. This measurement took the number of students who made it past the free trial, and divided it by the total number of students who enrolled in the class to begin with. This metric ended up being flawed in several reasons as an evaluation metric, which became evident during the calculations, thus why this metric was not actually used will be explained in the next two sections.

The number of users who enrolled in a course was also considered as an evaluation metric, but was not included for two reasons. First, we are unsure about how the sizing of users will be assigned to the control and experimental groups, so using a strict number count of users will make it hard to analyze our results. Will a difference in users simply be because of the different numbers in the assigned groups, or because there is a significant difference in the two experiences? This would be hard to determine using the number of user ids. Thus simply using a count will be much more difficult to evaluate than the gross and net conversion rates we currently have. Secondly because we have the gross and net counts, using user ids will also be a little repetitive, so there is no need to use it.

Launch Conditions:

In order to launch this experiment, the following two conditions must be met. The Gross Conversion rate must show a statistically and practical decrease in the experimental group and the Net Conversion rate must not show a statistical decreased measurement. If either of these two conditions are not met, I will recommend not launching this experiment because we would either be potentially losing customers or not reducing the number of frustrated students.

Measuring Standard Deviation

For all of our metrics, we are able to assume that our measurements will follow a Bernoulli distribution for the following two reasons. First we can assume all events are independent, and that one person signing up for a class will not affect another individual's actions. Also, Each event either has a success or failure action of clicking a button (or not), signing up for a class, and making it pass the free trial. Thus our baseline measurements (taken before the experiment) were used to come up with the following probabilities of success. Below are the baseline probabilities used and calculated standard deviation of each metric for a sample size of 5,000 cookies.

1. Gross Conversion: $p(enrollment \mid click) = 0.20625, \quad SD = 0.0202$
2. Net Conversion: $p(payment \mid click) = 0.1093125, \quad SD = 0.0156$
3. Retention: $p(payment \mid enrollment) = 0.20625, \quad SD = 0.0549$

The above standard deviations are all analytic estimates of the standard deviations for the evaluation metrics, which will work well for both our Net Conversion and Gross Conversion metrics. These two metrics' unit of analysis is the click on the "Start Free Trial" button, which so happens to also happens to be our unit of diversion. Thus we know that we have independent sampling, and that the analytical standard deviation will be a good estimate of our variability, and should be a close enough match to our empirical variations that we can use it to determine the sample size needed for this experiment.

Retention clearly now will not be a good evaluation metric for this experiment. Retention as a measurement is actually taking the number of students who pay and dividing it by the number of students who start the free trial. In order for an analytical variation to be a good estimate of variability, we would need to be able to assume independence of our unit of analysis (in this case the enrollment). However, it is possible that our experiment affects the number of students who enroll, thus we can not assume independence with retention, which will very much drive up our variability. If we were able to bootstrap, or conduct A/A tests, we could have better calculated the variability for retention using an empirical variation, but due to the limitations of this experiment (small data sample, no access to A/A tests), we are unable to do so. Right now with the high variability, we will need a much larger test (which we will show below), and this metric is quickly becoming less helpful than originally stated. Thus we will disregard retention as an evaluation metric, and more forward with just gross conversion and net conversion.

Sizing

Number of Samples vs. Power

A quick note before talking about the sizing of this experiment, whenever evaluating multiple metrics, even just two, it is worth considering using a multiple metric controlling procedure. However, for this experiment, we actually are evaluating two metrics separately, and therefore may need the Bonferroni correction to have more confidence in our results. For now though I will state that we will not be using the Bonferroni correction, and will better explain why in our results section below. Thus we will calculate the size needed for this experiment without such a correction.

Looking at our two metrics separately, we will calculate how many clicks we would need to determine statistical significance. Using an online sample size calculator, to determine a d_{min}= 0.0075, on a starting conversion rate of 0.109 at the Statistical power of $1-\beta=0.8$ and significance level $\alpha=0.05,$ Net Conversion we would need 27,413 clicks for each group. Using the same statistical power and significance power for Gross conversion, but with d_{min}= 0.01, and a starting conversion rate of 0.20625, you would need 25,835 clicks for each group to determine results of this metric. Because the Net conversion requires a larger number of clicks per group, we will use the first number determine our total number of page views needed:

27413 clicks for each sample x 2 (for control and experiment group) = 54,826 total clicks 54, 826 Total Clicks ÷ 0.08 (Probability of a click from a page view) = 685,325 Total pageviews.

Thus we need 685,325 Total Pageviews for this experiment to get large enough samples.

I quick note, again justifying why we did not use retention as an evaluation metric, we would have needed 4,741,212 of user ids per group to get a reliable calculation! That would require even more pageviews and clicks, and this sample size quickly becomes to large to manage. Clearly this number is way too high to be practical, and yet again, and for this reason it is impractical to use retention as.

Duration vs. Exposure

This experiment will be run with significant trade offs. To collect enough data in a reasonable number of days, we need to divert most of the traffic into this experiment. The website as is gets around 40,000 pageviews a day, but we need 685,325 pageviews. I want to make sure that information is collected for a couple of weeks, just to make sure there is no observable trends on weekends or weekdays that we can analyze. Determining that I want this experiment to run for 23 days, 75% of the website's traffic will have to be diverted to this experiment. This fraction however will be split in only about half seeing the experimental pop up, so about 37.5% of all traffic will be exposed to the experiment. This gives Udacity some risk (a little over a third of potential customers), but also does not make this experiment drag on for too long, so I believe it is a good trade off of potential risk to reward of making an informative improvement for the company.

Additionally, this experiment is a pretty low risk experiment in terms of the material and humans being exposed to the pop up. There is no additional information being collected, users are not being tracked individually, and there is definitely no sensitive information in the experiment. I doubt that there is any harm to come from this experiment, and the only things that are hopefully changing is the upfront understanding that the program will take time. Thus the experiment is not very risky in this regard, and I see no reason why we can not divert 75% of the traffic to the experiment.

Experiment Analysis

Sanity Checks

The two invariant metrics we wanted to check as sanity checks are number of cookies in each group and the number of clicks. After collecting data on 690,203 total page views and 56703 clicks, Here was our breakdown of observed values from our data for each group:

	Pageviews	Clicks	Click-Through Probability
Control	345,543	28,378	0.08212
Experiment	344,660	28,325	0.08218

We want to compare these count values for the first two metrics to what we theoretically should have seen in a 50-50 split. Below are the calculated confidence intervals of what percents of the number of clicks and page views we should have seen in our control group:

PageViews: (0.4988, 0.5012)

Clicks: (0.4959, 0.5041)

Our control group has 0.5006 of the pageview traffic (which was in the 95% confidence interval), and 0.5004 of the clicks (again in the confidence interval), so both of our sanity checks pass, and our experiment provided a fair split of our groups.
For the click through probability, we want to make sure that the differences in the probabilities are within the expected confidence intervals.

Click-Through Probability (difference) : (-0.0012, 0.0013)

Since the observed difference equals $d = 0.000056$, which is in our confidence interval, this sanity check also passes

Thus we can move onto the analysis of this test.

Result Analysis

Effect Size Tests

After looking at the test results, here are the following calculation values from this experiment. Note that the observed difference is in the probabilities of

Metric	Observed Difference (Exp. - Cont.)	95% Confidence Interval	Statistically Significant?	Practically Significant?
Gross Conversion	**-0.02055**	**(-0.0291, -0.0120)**	**Yes (Negative)**	**Yes (Negative)**
Net Conversion	**-0.00487**	**(-.0116, 0.0019)**	**No**	**Possibly**

It is worth noting here, that while the observed net conversion does not at first show practical significance, there is an issue worth noting which draws a cause for concern and why I wrote Possibly instead of no (which will be addressed in the recommendation section).

Sign Tests

On top of the above calculations, it was worth investigating if possible some single days of collecting data swayed the results one way or another. Using a less-robust sign test was also calculated just to see if there were discrepancies with these the previous results, simply to see if the results are worth investigating more. The sign test just counted the number of days we saw a positive difference, (where the experiment group's rate was higher than control group) and to see if was likely to occur if there was no statistical significant difference between the experimental and control groups. Here are these results:

Metric	Observed Positive Differences (Out of 23 Days)	P-value of Occurring (With 95% Confidence	Statistically Significant?
Gross Conversion	**4**	**0.0026**	**Yes**
Net Conversion	**10**	**0.6776**	**No**

Summary

If we had used Bonferroni's correction, we would have been more likely to have observed less false positives, at the expense of seeing more false negatives. We have a designed this test though where all of our metrics needed to land a certain way in order to recommend launching the change, so I think it was crucial to not potentially hide statistically significant results into false negatives, which we could be at risk for by using Bonferroni's correction. Also, since we are using two sided tests, I think it is crucial to make sure we alerted ourselves to any possibly negative side-effects, and we want to make sure that these are not being hidden in false negative results. Had we been solely looking for any positive statistical result I would have suggested to use Bonferroni's correction, however we weren't, and we wanted to make sure multiple metrics had certain outcomes.

Recommendation

From the above results, it seems as if the experiment met our desired results, and we should launch the experiment. However there is a crucial piece of information worth reviewing before launching the pop-up. The practical significance boundary of -0.0075 is within the 95% confidence interval of our observed net conversion rate. That means, while it is possible that the net conversion rate will not decrease significantly if we launch this change, it is just as possible that the experiment caused a decrease that is beyond our practical threshold. We just do not know exactly what our effect will be. That is, we may possibly be implementing a change that does in fact reduce our students who enroll past the free trial. For this reason then, I believe launching this change is too risky, and may cause us to lose business. I believe we should either not launch this change, or come up with a new idea to run through a follow up experiment that may better reduce frustrating students without potentially decreasing our net conversion rate.

Follow-Up Experiment

Now reflecting on the previous experiment, this experiment was well intentioned, but in my opinion has has a crucial issue.I believe the pop up was at the wrong time. A positive message to individuals who do sign up for a class would in my opinion help keep the net conversion up. I think maybe this pop-up could have the effect that Udacity was looking for the previous experiment.

In my proposed experiment, I would have a similar pop-up occur to students, however it would occur after they finished enrolling for a free trial. The pop-up would say:

"Congratulations on signing up for this course!
To make sure you are successful, we recommend spending 5 or more hours a week on this course!
If you need support, make sure to schedule a 1-on-1 appointment.
For resources on managing your time, click here."

The click here button would also provide a link to resources on time management, and possibly articles that show the importance of dedicating time to things. This message not only sends the clear expectation for students of the time commitment at the start of the enrollment, but it also does so in a positive way, while providing resources and support. This shows right away that Udacity will challenge students, demands a serious effort, but also will support them along the way. This pop up in my opinion would have a greater effect in improving student experiences in the free trial, thus making them more likely to stay on past the trial period, hence helping improve the net conversion rate. Thus my hypothesis for this follow-up experiment is that with this pop up, we can decrease frustrated students, while improving the net retention rate.

Because of the differences in this proposed experience, the following things would have to change. First, the unit of diversion should be the user ids instead of cookies. Since this pop up would only show up after individuals create an account, the user id would be the best way to split individuals into test/control groups. This also could be used as an invariant measurement, to judge if the control and test groups were roughly equally split in size.

The evaluation metric for this experiment could simply be retention. Since there should be no changes made until after students are enrolled, we would simply want to measure whether students are more likely to stay enrolled past the 14-day free trial period. In my opinion, sending this positive yet supportive message would help improve this metric, and hopefully it would do so in a significant way.

Resources

Online Sample Size Calculator:
http://www.evanmiller.org/ab-testing/sample-size.html

A/B Testing Support:
https://www.udacity.com/course/ab-testing--ud257