Designing an A/B Test: Free Trial Screener

Christopher Giler

2017-04-11

Experiment Design

Metric Choice

Invariant Metrics

1. Number of Cookies (Page Views): The number of unique cookies to view the course overview page. This metric is also the unit of diversion for this experiment, as we can ensure that the metric's population sizes are evenly split between the control and experiment groups. Because the number of cookies are evenly sized between the two test groups, it is not valid as an evaluation metric for the experiment.

2. Number of Clicks: The number of unique cookies to click the "Start free trial" button (which happens before the free trial screen is triggered). This metric is not affected by the experiment, as our interest is in the experiment's effect on who enrolls, rather than on who visits the page. For this reason, the number of clicks works as an invariant metric and not as an evaluation metric for the experiment.

3. Click-Through Probability: The number of unique cookies to click the "Start free trial" button divided by the number of unique cookies to view the course overview page. In other words, this metric is the number of clicks divided by the number of cookies or unique page views. Because both the numerator and denominator of this metric are invariant metrics, we can consider this to be an invariant metric as well, and not a valid evaluation metric.

Evaluation Metrics

1. Gross Conversion: The number of user-ids to complete checkout and enroll in the free trial divided by the number of unique cookies to click the "Start free trial" button. Because there is a possibility of the number of user-ids to complete checkout for the control and experiment groups, we cannot use this as an invariant metric. The gross conversion is also a credible way to quantify the probability of a user-id enrolling in a trial, so it is works well as an evaluation metric for this experiment.

2. Retention: The number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of user-ids to complete checkout. For this experiment, both the numerator and denominator of this metric can differ between the two test groups, so it does not work as an invariant metric. One result we wish to obtain from this experiment is whether or not clear expectations for the course workload is presented effectively before the student enrolls. Showing retention after a two-week time period is a good estimate of how effectively this information was explained to the student, so it makes for a good evaluation metric on this experiment.

3. Net Conversion: The 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. The net conversion is another metric which we could expect to see change between the experiment and control test groups, so it is not a valid invariant metric. One thing that we want to check is that showing the time commitment up front to prospective students does not negatively impact the number of students who enroll in the course past the trial period. In other words, we want to ensure that this change does not have a negative impact on revenue to Udacity. Therefore, net conversion works well as an evaluation metric to test this change.

Metrics Not Selected

1. Number of User IDs: The number of users who enroll in the free trial. This number cannot be distributed evenly between the two test groups in the experimental setup, and is therefore not an appropriate invariant metric. With cookies and clicks also selected as invariant metrics, it would be difficult to normalize both of these metrics in addition to the number of user-ids between the control and experiment groups, as page views and clicks do not require the user to be signed in. While the number of IDs is a good indicator of the number of students enrolling in a course, it does not give enough information alone to assess how the change helps to ensure students understand the requirements for the course prior to enrolling. For this reason, the raw number of user IDs is not as useful as an evaluation metric compared to the other metrics selected which provide a student conversion rate.

Desired Results for the Experiment

To consider the website change a success, the following results are needed in the evaluation metrics selected:

- **Gross conversion** for the experiment group is lower than that in the control group. This indicates that the change is effective in communicating to prospective students the time commitment needed to be successful in the course. For this change to be considered an improvement, it should reduce the enrollment of students who cannot make the time commitment required by the course. We expect this change to be both statistically and practically significant.
- **Retention** for the experiment group is higher than that in the control group. This would imply that the website change makes it known to users that enroll in the trial what the time commitment is to successfully complete the course prior to their enrollment. Retention can be viewed as the ratio of net conversion to gross conversion, and because we expect a decrease in gross conversion

with no negative impact on net conversion, this would require an increase in retention. We would expect this change to be both statistically and practically significant.

- **Net conversion** for the experiment group is not significantly lower than that in the control group. This demonstrates that the website change does not deter prospective students who are able to make the time commitment and remain enrolled in the paid course. Our desired result is that there is no practically or statistically significant decrease in this metric from the control to experiment groups.

Measuring Standard Deviation

To start, baseline data for each metric was reviewed to estimate the standard deviation analytically. The baseline data is given as:

Measure	Value
Unique cookies to view page per day	40,000
Unique cookies to click "Start free trial"	3,200
Enrollments per day	660
Click-through-probability on "Start free trial"	0.08
Probability of enrolling, given click (Gross Conversion)	0.2063
Probability of payment, given enroll (Retention)	0.53
Probability of payment, given click (Net Conversion)	0.1093

The standard deviation was estimated for a sample size of 5,000 cookies (or page views). Given this sample size, we can estimate 400 clicks on "Start free trial" and 82.5 enrollments. The standard deviation was calculated from this baseline data as follows:

Evaluation Metric	Sample Size	Probability	Standard Deviation
Gross Conversion	400	0.2063	0.0202
Retention	82.5	0.53	0.0549
Net Conversion	400	0.1093	0.0156

We can expect the analytical variance to match the emprical variance for Gross and Net conversion because the metrics are based on our unit of diversion for the experiment (number of cookies). Retention is based on number of user ids, which cannot be our unit of diversion for this experiment, so the analytical and empirical values are not expected to match.

Sizing

Number of Samples vs. Power

An estimate of page views required was calculated for each evaluation metric to determine feasibility of the required experimental setup. For these calculations, the Bonferroni correction was not applied (view summary section of the report for a detailed explanation). The number of page views were estimated for alpha = 0.05 and beta = 0.20.

Evaluation Metric	Minimum Detectable Effect	Baseline Conversion Rate	Number of Clicks	Number of Pageviews
Gross Conversion	1%	20.63%	25,835	645,875
Retention	1%	53%	379,297	**4,741,212**
Net Conversion	0.75%	10.93%	27,411	685,275

Running the experiment for these three evaluation metrics would require a total of 4,741,212 page views.

Duration vs. Exposure

The experiment would require directing half of website traffic to a control group (the current page), while the other half is directed to our proposed change. Because the change simply requires showing prospective students an additional page between clicking "Start free trial" and allowing the student to enroll, it would be reasonable to include 100% of traffic to the site for this experiment. This would require around 50% of traffic is exposed to the new change, and that no additional experiments are run for the duration of this experiment.

Estimating the amount of time required, assuming 40,000 page views a day, gives the following durations required for each evaluation metric.

Fraction of Traffic Exposed: **100%**

Metric	Duration
Gross Conversion	17 days
Retention*	119 days*
Net Conversion	**18 days**

To include retention as an evaluation metric for this experiment, the experiment would need to run for 119 days with 100% of traffic used for the experiment and no other experiments run simultaneously. Unfortunately, this timeline is unrealistic for the experiment, as we would like to see results much sooner than that to make a decision and/or run a follow-up experiment. For this reason, retention was removed as an evaluation metric at this stage. All following analyses and conclusions are based only the results for Gross Conversion and Net Conversion measurements.

Running this experiment poses low risk to both Udacity and Udacity customers. The experiment does not change the course content, website performance, or overall user experience. Because the experiment only looks at cookies, clicks, and enrollment/payment status of students, no sensitive information (such as student name or payment information) needs to be collected for this experiment to run.

Experiment Analysis

Sanity Checks

For each of of the invariant metrics selected, a 95% confidence interval was estimated for our expected observations. This was compared to the actual observed value for the metric, and comparing the value and acceptable range allowed us to confirm that our experiment was set up appropriately and that the invariant metrics were measured correctly.

Invariant Metric	Lower Bound	Upper Bound	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

Our sanity checks passed. All observed values for the invariant metrics fell within the expected range, as given by a 95% confidence interval.

Result Analysis

Effect Size Tests

For each evaluation metrics, a 95% confidence interval around the difference between the experiment and control groups was calculated. All calculations were performed without applying the Bonferroni correction.

Evaluation Metric	Lower Bound	Upper Bound	Statistically Significant?	Practically Significant?
Gross Conversion	-0.0291	-0.0120	Yes	Yes
Net Conversion	-0.0116	0.0019	No	No

Based on the experiment’s results, the difference for the gross conversion was found to be both statistically and practically significant, with the experimental group’s gross conversion rate being lower than that of the control group. There was no statistical or practical significance for the difference in the net conversion metric between the experiment and control groups.

Sign Tests

For each evaluation metrics, a sign test was conducted using the day-by-day data. The following table reports the p-value of the sign test and whether or not the result is statistically significant. All calculations were performed without applying the Bonferroni correction. The p-value was calculated for a 95% confidence level.

Evaluation Metric	p-value	Statistically Significant?
Gross Conversion	0.0026	Yes
Net Conversion	0.6776	No

The sign test confirms results from the effect size tests in the previous section. The decrease in gross conversion between the experiment and control groups was statistically significant. However, any change observed between the two test groups for net conversion was not statistically significant.

Summary

The Bonferroni correction is useful when we need at least one of the evaluation metrics to meet our expected outcome. However, in the case of this experiment, an improvement to the website would mean a statistically significant decrease in gross conversion rate as well as no statistically significant change in net conversion rate. Applying the Bonferroni correction to the experiment's data would mean the statistical analysis of our results is too conservative for what is required to make a decision.

The results of the experiment met expectations. A statistically and practically significant decrease in gross conversion rate was observed for the website change, while there was no significant change to the net conversion rate. In other words, while fewer students registered for a free trial, there was no significant impact on the number of students who allow their trial to roll over into a paid registration.

Recommendation

Taking the results of this experiment into consideration, I **do not recommend** carrying out the change to the website based on the data collected. This decision is due to the uncertainty in the results with respect to changes in net conversion. Gross conversion rate was effectively improved with the change. Giving prospective students a more personalized recommendation up front of the time required to be successful in the course led to a reduction in the number of students who enroll in a free trial. This means that a higher percentage of students in the trial are likely to continue on to a paid subscription, assuming there is no negative impact on net conversion as well.

However, this assumption cannot be made with confidence. Changes with respect to net conversion were found to be neither practically nor statistically significant. This means that we cannot postulate whether or not revenue to the company would be negatively impacted by this proposed change to the site. The confidence interval for the net conversion results is also biased negatively.

```
CI for Net Conversion : [-0.0116, 0.0019]
```

This negative bias suggests that there is a potential risk of net conversion decreasing if the change is launched. Before launching a change, it is important to ensure with statistical significance that this change will not have a negative impact on the business' net earnings.

Follow-Up Experiment

To follow up, I would recommend testing a potential improvement to user experience, with a focus on the 14-day trial period between enrollment and payment. To do this, an experiment could be run in which an additional project is added that the student can complete within the two week trial period, based on material which are considered prerequisites for the nanodegree. The course project could be presented as more of a tutorial or walkthrough, while still requiring some independent work by the student to complete. This type of walkthrough would give the student exposure to the nanodegree format as well as the project review process. A major advantage to enrolling in Udacity's Nanodegrees compared to other data science courses is the personalized feedback on course projects. Allowing the student to experience this advantage first-hand during the trial period could help improve retention.

For the experiment, we would want to look at retention as the evaluation metric. The invariant metric and unit of diversion would be user IDs, as this change would only impact those already signed up for a free trial. This experiment could also be performed alongside the change presented in this report, as this was found to successfully improve gross conversion. The change analyzed for this report should be applied to both the control and experimental test groups, and the follow-up change would be considered successful if the experiment showed a statistically and practically significant improvement in retention. This would suggest that launching this change would improve students' confidence in being able to complete future course projects successfully, and also leave a good first impression of Udacity's project review process.

In terms of looking at improvements in Udacity's profits, it could also be useful to quantify the cost of supplying this additional set of personalized project reviews. This cost can be compared to changes in revenue due to retention improvement, which would allow us to ensure that the increase in revenue also outweighs the operational costs associated with adding this new project to the curriculum.

References

- https://en.wikipedia.org/wiki/Bonferroni_correction
- http://www.evanmiller.org/ab-testing/sample-size.html (Calculating sample size)
- https://classroom.udacity.com/nanodegrees/nd002/parts/00213454013/project (Project description)