# A/B Testing Final Project

## Metric Choice

**INVARIANT METRICS**

There are several invariant metrics that could be used over the course of this experiment. In order to be fit, an invariant metric should not change across experimental and control groups. After conducting the experiment, these will provide a way to double-check the integrity of our design. Because the screener pops-up after clicking on the ’start free trial’ button, the number of pageviews, clicks, and the click-through-probability should remain unchanged during the experiment. Anything after the screener; number of user-id's, gross conversion, retention, and net conversion could be affected. Therefore, the invariant metrics chosen for this experiment are:

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

**EVALUATION METRICS**

Evaluation metrics are expected to change over the course of the experiment. By comparing differences between the control and experimental groups, we can measure the effect of the screener and test our hypothesis. Of the remaining metrics not considered to be invariant, user-id is excluded from the list of potential evaluation metrics. This is because user-id alone is a count, and gross conversion is a fraction that incorporates user-id while also offering a better way to track the effect of the screener. The evaluation metrics for this experiment are:

- 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)

If the hypothesis is correct, we would expect to see specific changes in the evaluation metrics. Gross conversion would be lower as those students likely to drop during the 14-day trial would be filtered by the screener. Retention rate would be higher as those likely to drop would not have enrolled, and those who enrolled would not be likely to drop. Last, net conversion would be unchanged as the amount of students to continue past the free trial and eventually complete the course would have been unaffected or there might be an increase in the net conversion because as the gross conversion rate decreases, we can say that the screener had an effect on enrollment thus resulting in higher net conversion.

## Measuring Variability

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:

5000/40000 = 0.125

Therefore, from 3,200 clicks and 660 enrollments in the baseline, 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 distrubution of the rates (gross conversion, retention, and net conversion) is gaussian. The standard deviation of these normally distributed rates is:

$$\sigma = \mathrm{sqrt}(p*(1-p)/n)$$

The rates for the evaluation metrics are:

p(gross conversion) = 0.20625
p(retention) = 0.53
p(net conversion) = 0.1093125

and calculation of the standard deviations yields:

$$\sigma(\text{gross conversion}) = \mathrm{sqrt}(0.20625(1-0.20625)/400) = 0.0202$$

$$\sigma(\text{retention}) = \mathrm{sqrt}(0.535(1-0.53)/82.5) = 0.0549$$

$$\sigma(\text{net conversion}) = \mathrm{sqrt}(0.1093125(1-0.1093125)/400) = 0.0156$$

Post experiment, the actual number of cookies used per day was higher than our estimated value. There were nearly 10,000 cookies per day per group rather than 5,000. Doubling the expected sample size for clicks to 800 and enrolls to 165, we can make a revised analytic estimate for standard deviation. The values that we expect to see are:

$\sigma$(gross conversion) = 0.0143
$\sigma$(retention) = 0.0388
$\sigma$(net conversion) = 0.0110

Of the 3 evaluation metrics, the analytically calculated standard deviation of retention is not likely to match the empirical standard deviation seen in the experiment. This is due to the fact that the units of diversion and analysis are different. The unit of analysis for retention is user-id, while the unit of diversion for the experiment is cookies. Gross conversion and net conversion have cookies as the unit of analysis, and the analytical standard deviation for these two metrics will likely match the empirical standard deviation seen in the experiment.

# Sizing

## Number of Samples vs. Power

I decide not to use Bonferroni correction, because the metrics in the test has high correlation and the Bonferroni correction will be too conservative to it.

I calculate the number of samples needed for each metric using the online calculator, with alpha = 0.05, 1 - beta = 0.2. The baseline conversion rate and minimum detectable effect (d_min) are listed individually below. Also note that the number produced by the online calculator is per branch, and in order to have both control and experiment, we need to double the number of required page views.

●Gross conversion. The baseline conversion rate is 0.20625, and d_min is 0.01. The required number of samples calculated from the online calculator is 25835. Note that this is the number of clicks on "start free trial", and in order to get that number, we need 25835 / 0.08 * 2 = 645875 page views.
●Retention. The baseline retention rate is 0.53, and d_min is 0.01. The required number of samples calculated from the online calculator is 39115. Note that this is the number of users who finished the 14 days free trial, and in order to get that number, we need 39115 / 0.08 / 0.20625 * 2 = 4741212 page views.
●Net conversion. The baseline conversion rate is 0.1093125, and d_min is 0.0075. The required number of samples calculated from the online calculator is 27413. Note that this is the number of clicks on "start free trial", and in order to get that number, we need 27413 / 0.08 * 2 = 685325 page views.
If we keep the retention rate as a evaluation metric, the number of required pages will be too large (in order to get 4.7 million page views, it takes 119 days of full site traffic, which is not realistic). Therefore we decide to drop the retention rate evaluation metric, and use gross conversion and net conversion as evaluation metrics, and the required number of page views (take the larger one) is 685325.

## Duration vs. Exposure

Considering the required pageviews, an exposure can be specified based upon the risk of the experiment, and from this a duration can be calculated. The exposure is dependent upon the risk involved and because the screener is a mild reminder about time commitment, it constitutes minimal risk. None of the participants could suffer physical harm as a result of the experiment, nor is sensitive data being collected, therefore a 100% exposure is a safe. 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 of an experiment and we should reduce the duration. We can exclude retention as an evaluation metric and consider the next limiting metric, net conversion. With a revised 685,275 necessary pageviews, it would then take 18 days to run the experiment.

Excluding retention as a metric still allows us to test our hypothesis with net conversion. The two metrics are highly correlated and reiterating the objectives will make this relationship clear. The intention of the screener is to reduce the amount of people who enroll that would otherwise drop-out during the 14-day trial, while unaffecting the number of people to go on past the 14-day trial. Retention measures the difference in the rate at which people drop from enroll to completion of the 14-day trial using user-id as both the unit of diversion and analysis. Net conversion also uses the number of user-id's to complete the trial as the unit of analysis and therefore captures the effect of retaining students. With a minimum detectable effect of 0.0075, the constraint on net conversion allows us to ensure that the same amount of students still go on to complete the trial.

# Experiment Analysis

## Sanity Checks

For counts ("number of cookies" and "number of clicks"), we model the assignment to control and experiment group as a Bernoulli distribution with probability 0.5. Therefore the standard deviation is `std = sqrt(0.5 * 0.5 / (N_1 + N_2))`, and the margin of error is `me = 1.96 * std`. The lower bound will be `0.5 - me` and the higher bound will be `0.5 + me`. The actual observed value is number of assignments to control group divide by the number of total assignments.

#### Number of cookies

control group total = 345543

experiment group total = 344660

standard deviation = sqrt(0.5 * 0.5 / (345543 + 344660)) = 0.0006018

margin of error = 1.96 * 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

The observed value is within the bounds, and therefore this invariant metric passed the sanity check.

#### Number of clicks on "start free trial"

control group total = 28378

experiment group total = 28325

standard deviation = sqrt(0.5 * 0.5 / (28378 + 28325)) = 0.0021

margin of error = 1.96 * 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

The observed value is within the bounds, and therefore this invariant metric passed the sanity check.

#### Click-through-probability on "start free trial"

For click through probability, we first compute the control value p_cnt, and then estimate the standard deviation using this value with experiment group's sample size, i.e. std = sqrt(p_cnt * (1 - p_cnt) / N_exp). The margin of error is 1.96 times of standard deviation.

control value = 0.0821258

standard deviation = sqrt(0.0821258 * (1-0.0821258) / 344660) = 0.000468

margin of error = 1.96 * 0.000468 = 0.00092

lower bound = 0.0821258 - 0.00092 = 0.0812

upper bound = 0.0821258 + 0.00092 = 0.0830

experiment value = 0.0821824

The observed value (experiment value) is within the bounds, and therefore this invariant metric passed the sanity check.

## Effect Size Tests

Let N denote the number of total samples (denominator) and X denote the number of target samples (numerator), and _cnt denote controlled group and _exp the experiment group. We first computed pooled probability and pooled standard error as

p_pooled = (X_cnt + X_exp) / (N_cnt + N_exp)

se_pooled = sqrt(p_pooled * (1-p_pooled) * (1./N_cnt + 1./N_exp))

The probability difference is computed as

d = X_exp / N_exp - X_cnt / N_cnt

With these values in hand, the lower bound and upper bound are

lower = d - se_pooled
upper = d + se_pooled

### Gross conversion

For gross conversion, the total samples (denominator) are the clicks of "start free trial", and the target samples (numerator) are enrolled users. The caculation is shown below.

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 * (1-p_pooled) * (1./N_cnt + 1./N_exp)) = 0.00437

d = X_exp / N_exp - X_cnt / N_cnt = -0.02055

lower = d - se_pooled = -0.0291
upper = d - se_pooled = -0.0120

Since the interval does not contain 0, the metric is statistical significant. It does not include d_min = 0.01 or -d_min = -0.01either, and therefore it is also practical significant.

### Net conversion

For net conversion, the total samples (denominator) are the clicks of "start free trial", and the target samples (numerator) are paid users. The calculation is shown below.

N_cnt = clicks_controlled = 17293.
X_cnt = pay_controlled = 2033.
N_exp = enroll_experiment = 17260.
X_exp = pay_experiment = 1945.

p_pooled = (X_cnt + X_exp) / (N_cnt + N_exp) = 0.1151
se_pooled = sqrt(p_pooled * (1-p_pooled) * (1./N_cnt + 1./N_exp)) = 0.00343

d = X_exp / N_exp - X_cnt / N_cnt = -0.0048 lower = d - se_pooled = -0.0116

upper = d + se_pooled = 0.0019

Since the interval contains 0, it is not statistical significant, and consequently not practical significant either.

## Sign Tests

We use the online calculator to perform sign test.

●For gross conversion, the number of days we see an improvement in experiment group is 4, out of total 23 days of experiment. With probability 0.5 (for sign test), the online calculator calculates a p-value 0.0026, which is smaller than alpha = 0.05. Therefore the change is statistical significant.

●For net conversion, the number of days we see an improvement in experiment group is 10, out of total 23 days of experiment. With probability 0.5 (for sign test), the online calculator calculates a p-value 0.6776, which is larger than alpha = 0.05. Therefore the change is not statistical significant.

## Summary

The effect size tests determine that gross conversion is both statistically and practically significant, while net conversion is neither. The gross conversion rate dropped in the experimental group by approximately 2% and thus the screener proved to be effective at reducing the number of students that enrolled from initial click. This supports our hypothesis if the screener is to be effective. Net conversion, however, was reduced by approximately 0.5% indicating that the screener had a negative effect on the number of students that would go on to complete the 14-day trial. In other words, the screener deterred some students from enrolling that would have otherwise completed the trial. This is not our intended effect and does not support the hypothesis.

The Bonferroni correction was not used in the analysis phase because our launch decision is based upon the significance of two metrics rather than just one. Had we used just one metric out of several to base our launch decision, the Bonferroni method would be appropriate. But, because the nature of our hypothesis requires that two effects be considered, we cannot base our decision in one metric alone.

The sign tests allow for an additional form of analysis. The conclusion from the sign test mirrors that of the effect size test, that gross conversion is significant but net conversion is not. Had we any discrepancies with regard to the significance of the evaluation metrics between the sign and effect size tests, further study would be warranted. In this case, both tests agree and our conclusions with regard to both metrics are strongly supported.

## Recommendation

The screener proved to be effective at reducing the number of people to continue from click to enroll, but it was not successful an unaffecting the number of students that would continue on past the 14-day trial. In fact, the screener appeared to increase the rate at which people left the 14-day trial. Based upon this evidence, we should not launch the change to Udacity’s website.

## Follow-Up Experiment

A follow-up experiment could be based upon motivation with only a slight change from the previous experiment. It would require a method to approximate the number of hours that each student dedicated to the material in the first week. If a student committed less than the recommended number of hours, a message would pop-up upon login before the start of the second week to motivate the student to commit more time. Along with this message could be links to success stories or video of interviews with previous students who have completed that particular class or nanodegree. It could include information about where they went on to land a new job with their newfound skills from that class or nanodegree. Those that met the recommended number of hours wouldn’t get a direct message but could still access this repository of interviews from their course homepage under the ’Resources’ tab.

My hypothesis is that the message would motivate some students who might otherwise drop out during the 14-day trial to continue past and possibly complete the course. It would also not affect those people that would otherwise continue through the trial and complete the course had there been no pop-up message. In this case, the overall student experience in the forums could be more energized and improve beyond the first week, and coaching resources would be used on more enthusiastic and dedicated students.

Considering this design, retention rate would be the best way to test our hypothesis. Being that we would divert traffic evenly among the control and experimental groups, the most suitable invariant metric would be the number of user-id's to complete checkout and enroll in the course. So we could consider user ID as a great choice for unit of diversion. It would be practical at this point, to divert students into the control or experiment group. It’s worth noting that this experiment would likely take longer to conduct, recalling that the inclusion of retention rate added significantly more time to the duration than gross or net conversion.