Free Trial Screener A/B Test

Design an A/B Test to Decide Whether to Launch a Free Trial Screener

By George Liu

A/B test is an effective and powerful tool in web site optimization and app development. It is widely used by data scientists in the technology industry. In this project, we will go over the full process of A/B testing and design an A/B test for Udacity in order to make a decision about the launch of a Free Trial Screener.

The Screener is a popup window advising students of recommended weekly time commitment, and based on the response, suggesting the user either to proceed with registration or to continue evaluating courses for free. The goal of using the Free Trial Screener is to reduce churn rate during the free trial period, and to improve student experience plus Udacity's coaching support capacity.

The hypothesis is 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.

Experiment Design

1. Metric Choice

In this project, the unit of diversion is cookie and we have seven potential metrics that can be used for either invariant or evaluation metrics. The candidates are:

- **Number of cookies**: number of unique cookies to view the course overview page.
- **Number of user-ids**: number of users who enroll in the free trial.
- **Number of clicks**: number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is triggered).
- **Click-through-probability**: number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.
- **Gross conversion**: 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.
- **Retention**: number of user-ids to remain enrolled past the 14-day boundary (an thus make at least one payment) divided by number of user-ids to complete checkout.
- **Net conversion**: 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 following table gives the metric choices and corresponding reasoning:

Metrics	Invariant	Evaluation	Reasoning
Number of cookies	Yes		This is the unit of diversion, so should evenly split across both groups, therefore good for invariant checking. In the meantime, the invariant characteristic means that it cannot represent the free trial screener's effect difference on both groups, thus not good for evaluation purposes.
Number of user-ids			This is the very indicator our free trial screener is supposed to change, thus not good for invariant checking. On the other hand, as this is an absolute number, rather than a ratio, it is not ideal for evaluation either.
Number of clicks	Yes		Since the clicks happen before the free trial screener is triggered, this metric should be roughly the same among two groups, so it's a good invariant metric. As such, it is not good for measuring performance as an evaluation metric.
Click-through-probability	Yes		This is the combination of number of cookies and number of clicks, and as both are good invariant metrics, the click-through-probability is one too. The invariant characteristic means it's not good for evaluation purposes.
Gross conversion		Yes	When users see the screener, it's likely that some will decide to think again and choose evaluate for free, so the gross conversion rate should drop. Therefore, it's not invariant, and is a good indicator of how many few people will enrol, i.e. an evaluation metric.
Retention			Originally, this metric was chosen as an evaluation metric, but was later dropped. It was chosen as it represents the potential effect of retaining more users by the free trial screener. However, later calculations indicate that a very large sample size is required to use this metric, therefore, it was dropped and replaced with net conversion.
Net conversion		Yes	Similar to gross conversion, this metric measures how the percentage of people who click free trial button ultimately make the payment. This metric should remain roughly the same because the percentage of people who have enough time commitment shouldn't change as a result of the screener. It could also drop, but definitely not too much as that'll negatively affect business results. As such, net conversion is not an invariant metric, but is a good evaluation metric.

2. Measuring Standard Deviation

Next, we'll make an analytic estimate of the evaluation metrics' standard deviation, given a sample size of 5000 cookies visiting the course overview page.

Given data:

Unique cookies to view page per day:	40000
Unique cookies to click "Start free trial" per day:	3200
Enrollments per day:	660
Click-through-probability on "Start free trial":	0.08
Probability of enrolling, given click:	0.20625
Probability of payment, given enroll:	0.53
Probability of payment, given click	0.1093125

Based on this, for the actual sample, we have the following:

Unique cookies to view page per day:	5000
Unique cookies to click "Start free trial" per day:	5000 * 0.08 = 400
Enrollments per day:	400 * 0.20625 = 82.5
Click-through-probability on "Start free trial":	0.08
Probability of enrolling, given click:	0.20625
Probability of payment, given enroll:	0.53
Probability of payment, given click	0.1093125
Payments per day	82.5 * 0.53 = 43.725

Now we'll calculate the actual proportions and the corresponding standard deviations with above data:

Proportions	Value	SE
Gross Conversion	= enrolments / clicks = 82.5 / 400 = 0.20625	= $\sqrt{p * (1 - p) / n}$ = $\sqrt{(0.20625 * (1 - 0.20625))/400}$ = 0.0202
Retention	= payments / enrolments = 43.725 / 82.5 = 0.53	= $\sqrt{p * (1 - p) / n}$ = $\sqrt{(0.53 * (1 - 0.53))/82.5}$ = 0.0549
Net Conversion	= payments / clicks = 43.725 / 400 = 0.1093125	= $\sqrt{p * (1 - p) / n}$ = $\sqrt{(0.1093125 * (1 - 0.1093125))/400}$ = 0.0156

If we were to calculate the standard errors empirically, the values for retention would have been much higher than the analytically calculated values, while there should be close values for both gross conversion and net conversion values. This is due to the fact that the unit of analysis (denominator) for retention is the number of user-ids which is different from the unit of diversion (cookies). On the other hand, gross and net conversion both have cookies as their unit of analysis which is same as the unit of diversion.

3. SIZING

- Number of Samples vs. Power

  In this test, we'll use alpha = 0.05 and beta = 0.2. In order to get the number of page views needed, we need to calculate the sample sizes for each of the three evaluation metrics, and add them together to get the result. Using Evan's Awesome A/B Tools, we get the following sample sizes:

  - Gross Conversion: 25,835
  - Retention: 39,115
  - Net Conversion: 27,413

  However, these are the sample sizes for the unit of analysis (denominator). We can use corresponding scaling factors to calculate the page views needed:

Metrics	Page Views Needed
Gross Conversion	= (Sample Size / Click-through-probability) * 2 = (25835 / 0.08) * 2 = 645875
Retention	= (Sample Size / (Click-through-probability * Probability of enrolling given click)) * 2 = (39115 / (0.08 * 0.20625)) * 2 = 4741212
Net Conversion	= (Sample Size / Click-through-probability) * 2 = (27413 / 0.08) * 2 = 685325
Total	6,072,412

  Since this test involves multiple metrics, I decided to use Bonferroni correction (this decision was later changed. Please refer to summary for detailed explanation). The total number of page views needed is 6,072,412.

- Duration vs. Exposure

  We'll start by evaluating the risks associated with this experiment. There are two major types of risks - consumer and business risks. The former one represents any risks a consumer can encounter when exposed to the experiment. In this case, it's minimal risk and thus there is no safety concerns for us to limit the exposure. As for the business, although adding the free trial screener may lead to decreased net conversion, possibly translating to lower revenue from student payments for the business, even if we limit exposure, over time, the potential number of payed membership loss is still the same. Therefore, there's no need to limit exposure either. So all the traffic will be diverted to this test given Udacity is not running any other tests at the moment. If we use the above calculated total number of page views, we can calculate the length of experiment as:

Days Needed = Number of Page Views Needed / (Available Traffic * 50%)

= 6072412 / (40000 * 100%) = 151.8

Of course, no business is willing to wait for half a year to decide on a page optimization decision. Therefore, I decided to drop the retention metric, and also to use the same traffic for both gross conversion and net conversion. This way, I only need to select the bigger of the two sample sizes to calculate the length of experiment:

Days Needed = Number of Page Views Needed / (Available Traffic * 100%)

= 685325/ (40000 * 100%) = 17.1 ≈ 18 days

So, we need about 3 weeks to carry out the test with all our traffic exposed.

Experiment Analysis

4. Sanity Checks

Before any further analysis, we first need to do sanity checks to make sure that the data we get are not "corrupted" in any way, for example, maybe due to experiment setup or infrastructure issue, we get different numbers of page views in each group. The sanity check can be done by checking the invariant metrics.

Below, for each invariant metric, we'll compute a 95% confidence interval for the value that's expected. The observed value is also given. If observed value is between the confidence intervals, the metric passes the sanity check.

For each metric, since each assignment of a cookie is a random event with two outcomes, i.e. either be assigned to the control or the experiment group, we can then use binomial distribution with a probability of 0.5 to estimate the expected number of cases in each group.

	Cookies (Control / Experiment)	**Clicks (Control / Experiment)**	**Click-through-probability (Control / Experiment)**
Count	345543 / 344660	28378 / 28325	0.0821 / 0.0822
Percentage or Difference	0.5006	0.5005	0.0001 (difference) $p_{pool} = \frac{\#Clicks_{ctr} + \#Clicks_{exp}}{\#Cookies_{ctr} + \#Cookies_{exp}} = 0.08215$
SE	$=\sqrt{\frac{0.5 * (1-0.5)}{345543 + 344660}}$ = 0.000601841	$=\sqrt{\frac{0.5 * (1-0.5)}{28378 + 28325}}$ = 0.002099747	$=\sqrt{p_{pool} * (1 - p_{pool}) * (\frac{1}{\#Cookies_{ctr}} + \frac{1}{\#Cookies_{exp}})}$ = 0.000661061
Upper Bound	0.5012	0.5041	0.0013
Lower Bound	0.4988	0.4959	-0.0013
Observed Value	0.5006	0.5005	0.0001
Pass	Yes	Yes	Yes

So all three invariants pass the sanity check as the observed values all fall between the expected intervals. Since the total numbers pass the check, there is no need to investigate the day by day numbers. We can proceed with the result analysis.

5. Result Analysis

- Effect Size Tests

  For each of the evaluation metrics, we'll compute a confidence interval around the difference. Also, I've decided not to use the Bonferroni correction since we're using "AND" on all evaluation metrics instead of "OR" as per the hypothesis.

	Gross Conversion	**Net Conversion**
$\hat{d}$ **(Sample Diff)**	$= \text{Gross Conversion}_{exp} - \text{Gross Conversion}_{ctr}$ = 0.198319815 - 0.218874689 = -0.020554875	$= \text{Net Conversion}_{exp} - \text{Net Conversion}_{ctr}$ = 0.112688297 - 0.117562019 = -0.004873723
p_{pool}	$= \frac{Enrollments_{ctr} + Enrollments_{exp}}{Clicks_{ctr} + Clicks_{exp}}$ $= \frac{3785 + 3423}{17293+17260}$ = 0.208607067	$= \frac{Payments_{ctr} + Payments_{exp}}{Clicks_{ctr} + Clicks_{exp}}$ $= \frac{2033 + 1945}{17293+17260}$ = 0.1151275
SE	$= \sqrt{p_{pool} * (1 - p_{pool}) * (\frac{1}{\#Clicks_{ctr}} + \frac{1}{\#Clicks_{exp}})}$ = 0.004371675	$= \sqrt{p_{pool} * (1 - p_{pool}) * (\frac{1}{\#Clicks_{ctr}} + \frac{1}{\#Clicks_{exp}})}$ = 0.003434134
Upper Bound	-0.0120	0.0019
Lower Bound	-0.0291	-0.0116
Statistical Significance Boundary	0	0
Statistical Significant	Yes	No
Practical Significance Boundary	0.01	0.0075
Practical Significant	Yes	No

Note:

- *Since only first twenty-three days' data include enrollments info, we use this period for calculation instead of the full date range.*
- *A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)*

- Sign Tests

  Now we'll conduct a sign test as a verification for the previous effect size test. To do this, we'll calculate the percentage of expected change for both evaluation metrics and use the online calculator to get the p-value for this to happen randomly. The expected change is

calculated as difference between experiment value and control value as shown in the following image (highlighted cells).

Gross Conversion Diff	Net Conversion Diff
-0.041989722	-0.052329603
-0.040932765	0.026064774
-0.019691223	-0.015143935
-0.019734673	-0.014352621
-0.026473899	0.036517209
-0.003973639	-0.022224312
-0.032366653	-0.045194022
-0.029878854	-0.015667469
-0.017413891	0.023642778
-0.013730654	0.00129379
-0.060557638	-0.038931341
-0.033517173	-0.022394441
-0.000946291	0.021708477
-0.048558778	-0.046414159
-0.064867753	-0.064162551
-0.006621909	-0.00114951
-0.030718281	-0.009027853
0.010869565	0.039402174
0.011255405	0.015676041
0.056820223	-0.010146869
0.005618639	0.027300621
-0.024758343	0.007321015
-0.045877082	0.045584097

Based on the above data, we have the following:

	Gross Conversion	**Net Conversion**
Expected Change	Decrease	Decrease
Count of Expected Change Cases	19	13
Total Cases	23	23
Two-tail P-value	0.0026	0.6776
Alpha	0.05	0.05
Significant	Yes	No

This result perfectly aligns with our conclusion from previous effect size test result.

- Summary

  Based on our test results, the free trial screener has a significant effect on gross conversion, both statistically and practically. It lowers the gross conversion rate as expected,

since there should be less enrollments due to the suggestion of time commitment provided by the screener.

On the other hand, the screener doesn't have a significant effect on net conversion which is a preferred result since net conversion represents business performance and revenue.

I decided not to use Bonferroni correction. This is due to the fact that our hypothesis states that both gross and net conversion conditions should be met, this is an "AND" situation which already imposes a conservative requirement for us to get rid of false positives, therefore, there is no need to use Bonferroni correction to be conservative again. However, if an "OR" situation exists between the evaluation metrics, we will definitely use the Bonferroni correction.

6. Recommendation

From previous analysis, we have:

Gross Conversion confidence interval: -0.0291, -0.0120 (practical boundary 0.01)

Net Conversion confidence interval: -0.0116, 0.0019 (practical boundary 0.0075)

Therefore, our results show that the trial screener will reduce the number of risky enrolments (gross conversion) who might cancel due to time constraint – this is what we expect and want. However, since the net conversion confidence interval does include the negative practical boundary, it is then likely that net conversion could drop at a level that the business cares. This contradicts our initial hypothesis of reducing gross conversion while maintaining the percentage of paying students. As a result, I would recommend not to launch the free trial screener given the negative revenue impact.

Follow-Up Experiment

In order to further reduce the number of frustrated students who cancel early in the course, we can conduct a follow-up experiment as follows:

Design a message that will display when a user clicks the cancelation button early in the course. This message can include the benefits of a payed membership, such as coaching support etc. The message will be displayed for the experiment group and will be unavailable for the control group. The hypothesis is that when being presented the benefits of a payed membership, a user will reconsider the cancelation decision and thus the message may reduce the probability of cancelation.

In order to measure this, we can select the number of user-ids who cancel early divided by the total number of user-ids as the evaluation metric. As per the hypothesis, this proportion should drop in the experiment group. To implement, we can divert our traffic to control and experiment groups by using user-id as the unit of diversion, so that some users will see the message and the rest won't.

Reference

1. Evan's Awesome A/B Tools
2. Sign and binomial test
3. When to use Bonferroni correction