A/B Test to Decrease Early Dropout of Free Trial

By Bill Kapsalis

Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.
In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead.This screenshot shows what the experiment looks like.
The hypothesis was 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. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.(1)

Experiment Design

Metric Choice

Invariant Metrics

Number of cookies, Number of clicks and Click-through-probability were used as invariant metrics because these happens before the free trial screener is trigger and would not be affected by the experiment.

Number of cookies: That is, number of unique cookies to view the course overview page.(1)

Number of clicks: That is, number of unique cookies to click the "Start free trial" button.(1)

> **Click-through-probability:** That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.

These invariant metrics should not change across the control and experiment groups. These metrics are formed before the free trial screen is triggered. The cookie is the unit of diversion. Every time a cookie is formed it is randomly placed in the control group or the experimental group. Therefore there should not be a statistically significant difference in the count(or other factors) of these metrics. If there is a difference the experiment may not be designed properly.

Evaluation Metrics

The high level idea here is the first part of the hypothesis, that is ‘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’. This can be evaluated with the evaluation metric ‘Gross conversion’.

> **Gross conversion:** That is, 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.(1)

The Gross conversion shows the effect of the experiment on whether or not a person enrolls in the free trial after clicking the enroll button. After seeing the experimental message “Udacity courses usually require a greater time commitment for successful completion” the person may think “I can only put in 3 hours a week so I will not enroll”. This would cause the Gross conversion ratio to decrease. This would not be a bad thing if the message also decreased the number of students that leave the free trial.

One metric is not enough because the hypothesis has two parts.

The second part of the hypothesis is ‘without significantly reducing the number of students to continue past the free trial and eventually complete the course.’ To determine this we can use the metric ‘Net conversion’.

> **Net conversion:** That is, 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. (1)

If there is no statistical significant decrease of this metric in the experimental group this would tell us the number of students that finished the free trial and made at least one payment was not changed by the experimental message.

Unused metrics

> **Number of user-ids:** That is, number of users who enroll in the free trial.

This metric was not used as an invariant metric because it is formed after the free trial screen is triggered. Also, it is not directly used as an evaluation metric because it is already used as the numerator of the Gross conversion variable that was used.

> **Retention:** That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout.

The Retention metric was not used as an invariant metric because it is formed after the free trial screen is triggered. This was initially evaluated as an evaluation metric but it had to much variation. To get the power needed for this experiment using this metric would require an unreasonable number of samples to finish this experiment in a reasonable amount of time.

I will be looking for a decrease in Gross conversion and no decrease in Net conversion. A decrease in Gross conversion will show that fewer people enrolled because we know the denominator 'Number of cookies', the invariant metric, did not change. Also, I will be looking for no decrease in Net conversion. This will show that the experimental message had no effect on the number of students finishing the free trial. Since we know fewer people enrolled and the same number finished the free trial we can conclude that fewer people left the free trial. Therefor if the Gross conversion decreased and the Net conversion stays the same we can conclude the hypothesis is true and we should launch the change.

Measuring Standard Deviation

Analytical Standard Deviation:
Gross conversion has a standard deviation of 0.0202.
Retention has a standard deviation of 0.0549
Net conversion has a standard deviation of 0.0156.

Dr Tang said(2) that the analytical variance for some of the simple metrics are an underestimation of the variance. Therefore I did an empirical calculation of the variance for gross conversion, retention and net conversion to see if this is the case here. I created a difference between the control and experimental groups for the 3 metrics. Then I took the standard deviation.

Empirical Standard Deviation:
Gross conversion has a standard deviation of 0.0270.
Retention has a standard deviation of 0.1379.
Net conversion has a standard deviation of 0.0313.

The analytical variance can be an underestimate of the variance if the unit of analysis is different than the unit of diversion.(2) This is not the case with Gross conversion and Net conversion so for this discussion I will look at the Retention metric. The unit of analysis is the denominator of the metric. For the Retention metric this is ‘user-ids to complete checkout’ and this is different from the unit of diversion which is unique cookies. The analytic variance did underestimate the variance. Here the empirical variance is 2.51 times the analytic variance.

Here the unit of analysis and the unit of diversion are both unique cookies for Gross and Net conversions. Therefore the slight underestimation of the analytical variances is probably due to the lack of independence of events caused by the unique cookie for ‘groups of events’.(2)

Sizing

Number of Samples vs. Power

We need to determine how many unique cookies we need to get a statistically significant result. This is statistical power. We need enough power to determine with a high probability that our observations are statistically significant. To detect a smaller change or a higher confidence level we need more power, that is more unique cookies(pageviews).

This online calculator(3) was used to determine the number of pageviews for an alpha= 0.05 and beta=0.2.

Evaluation metric Gross conversion:
Number(from calculator) of ‘user-ids to complete checkout and enroll’ needed for power:
25835
Conversion factor to get ‘unique cookies to click the "Start free trial" button’ needed for power:
40000/3200 = 12.5
Number of ‘unique cookies to click the "Start free trial" button’ needed for power:
25835*12.5 = 322937.5
Multiplied by 2 for control and experimental groups:
322.937.5 * 2 = 645875

Evaluation metric Retention:
Number(from calculator) of ‘user-ids to remain enrolled past the 14-day boundary’ needed for power:
39115
Conversion factor to get the number of ‘user-ids to complete checkout’ needed for power:
40000/660 = 60.6
Number of ‘user-ids to complete checkout’ needed for power:
39115*60.6 = 2370606.1
Multiplied by 2 for control and experimental group:

2370606.1 * 2 = 4741212.2

Evaluation metric Net conversion:
Number(from calculator) of ‘number of user-ids to remain enrolled past the 14-day’ needed for power:
27413
Conversion factor to get the number of ‘unique cookies to click the "Start free trial" button’ needed for power:
40000/3200=12.5
Number of ‘unique cookies to click the "Start free trial" button’ needed for power:
27413*12.5 = 2370606.061
Multiplied by 2 for control and experimental group:
2370606.061 * 2 = 645875

The Retention metric required too many cookies to compete the experiment in a reasonable amount of time. Therefore the Retention metric was not used here or for effect size test.

The next highest number is for the metric Net conversion. This has 685,325 unique cookies(pageviews) needed for the control and experimental group to generate the needed power.
I did not use Bonferroni correction because it is overly conservative.

Duration vs. Exposure

I did not want to jeopardize all the pageviews in case there was an unseen problem. There may be a problem with the experimental user interface feature in some browsers(8). This may not allow any people on certain browsers to enroll. This would frustrate the person and cost a lot of money if the frustrated people just stop trying to enroll. Alternatively, press or a blog(8) is more likely to see and possibly comment on a change that may not be permanent. For these reasons I exposed 50% of the daily page views to the experiment and control groups for a duration of 35 days.

Experiment Analysis

Sanity Checks

Number of cookies 95% confidence level:
Upper bound: 0.5012
Lower bound: 0.4988
Observed: 0.5006
Passes: TRUE

Number of clicks on “Start free trial” 95% confidence level:

Upper bound: 0.5041
Lower bound: 0.4959
Observed: 0.5005
Passes: TRUE

Click-through-probability on “Start free trial” 95% confidence level:

Upper bound: 0.0013
Lower bound: -0.0013
Observed: 0.0001
Passes: TRUE

These invariant metrics did not change across the control and experiment groups.

Result Analysis

Effect Size Tests

Gross Conversion 95% confidence level:

Upper bound: -0.0120
Lower bound: -0.0291
Statistically significant: TRUE
d_min= 0.01
Practically significant: TRUE in a negative direction.

Net Conversion 95% confidence level:

Upper bound: 0.0019
Lower bound: -0.0116
Statistically significant: We do not have enough power to draw a strong conclusion. At this power level zero is within the CI so technically there is not a significant statistical change.
d_min= 0.0075
Practically significant: We do not have enough power to draw a strong conclusion. At this power level negative d_min(-0.0075) is within the CI so technically there is not a significant practical change.

Sign Tests

Gross conversion: Out of 23 days of results only 4 showed a positive difference in Gross conversion.
The p-value of the sign test is: 0.0026
Statistically significant negative difference: TRUE

Net conversion: Out of 23 days of results 10 showed a positive difference in Net conversion.
The p-value of the sign test is: 0.6776
Statistically significant difference: FALSE

Summary

I did not use Bonferroni correction because it would decrease the power and is overly conservative.The Bonferroni correction makes no assumptions about independence or dependence of the multiple metrics being tested simultaneously(9). One problem here is our data is correlated and therefore Bonferroni correction would be overly conservative and we may miss statistically significant changes.

Here we have two aspects of the hypostasis that must be met. These are a ‘decrease’ of the Gross conversion metric and ‘no decrease’ of Net conversion metric. Both of these criteria must be true to trigger the launch of the changes. Put another way ‘ALL’ of the criteria must be true to launch.

Bonferroni correction adjusts the Pvalue to minimize the chance of a false positive (type 1 errors) of one of the many metrics being tested simultaneously. In an experiment where a false positive of ‘ANY’ criteria would result in an inappropriate launch of changes Bonferroni correction would be justified. However, in our experiment ‘ALL’ of the criteria must be met to launch the changes. Since our experiment has two aspects the chance of getting a false positive in both , resulting in an inappropriate launch, is much less. Additionally, decreasing alpha, as in Bonferroni correction, would decrease the power of the experiment. This increases the chances of a false negative (type 2 errors). This would decrease our ability to detect smaller differences between the control and experimental groups. This could cause us not to launch the changes appropriately. Therefore I did not use Bonferroni correction.

The ‘sign test’ did match the ‘effect size test’ result for Gross conversion. Both had a significantly statistical decreased at a 95% confidence level. The Net conversion sign test did not show a significant change at a 95% confidence level. The Net conversion ‘effect size test’ was not conclusive. The reason the sign test clearly produced ‘no significant change’ and the effect size test did not may be because the number of samples was to low. Maybe the power needs to be increased by increasing the number of samples. Another reason for the difference between the sign test and the size effect is ‘Net conversion by day’ shows a wide range.
Net conversion ‘by day of the week’:

- Saturday: 0.1025
- Sunday: 0.1283
- Monday: 0.1177
- Tuesday: 0.1166
- Wednesday: 0.1102
- Thursday: 0.0908
- Friday: 0.1187

The low of the range is 0.0908 and the high of the range is 0.1283. This variation by day may signify a Simpson's paradox. Simpson's paradox is when the subgroups have stable results but the mix of the subgroups, as in the 'effect size test', may have different results. This may explain why the sign test above shows 'no decrease' and the size effect for Net conversion shows inconclusive results.

The Net conversion confidence interval includes zero and the negative d_min(0.0075). The CI barely includes zero, by 0.0019 with a CI range of 0.0135. Since both are included in the CI a statistical and practical significant decrease in Net conversion can not be determined at this power level.

Recommendation

The Gross conversion metric has a statistically and practically significant decrease at a 95% confidence limit but the Net conversion metric has a wide CI that includes zero and the negative d_min(0.0075) so we do not have enough power to draw a strong conclusion. Run an additional test with higher power is recommended. Since the SE is inversely proportional to the number of samples I recommend running the test again twice as long. If there is no time, money or desire to rerun the test I recommend not implementing the changes because there is a possibility of a practically significant decrease in Net conversion. This would indicate a reduction in enrollment past the free trial and decreased payments.

After Rerunning with More Samples(Higher Power)

This may result in a narrower CI range. This new range may:

- Still include zero and the negative d_min(-0.0075):
  - This would need a judgment call to implement changes. This may or may not significantly reduce the number of students to continue past the free trial and make at least one payment.
- Be narrower and include only zero:
  - The two aspects of the hypothesis were proven true. First let's look at the second aspect which is "...without significantly reducing the number of students to continue past the free trial". The Net conversion metrics shows no significant change. If Net conversion has no significant change this aspect of the hypothesis would be true. So, at this power there is no statistically significant change to Net conversion.

    We need to look at both the Gross conversion and the Net conversion to see if the first aspect of the hypothesis is meet. Both metrics have the same denominator, that is 'unique cookies to click the "Start free trial" button.' The first aspect of the hypothesis is setting clearer expectations for students upfront would reduce the number who left the free trial. We know the denominator of both metrics is the same. Also there was no statistically significant change to the Net conversion metric. Therefor we can conclude there was a decrease in

enrollments into the free trial. Additionally, since there was no decrease to Net conversion we can conclude there was a ‘reduction’ in the number who left the free trial. Therefor implement the changes.

Be narrower and is between zero and negative d_min(-0.0075):

We would have a statistical decrease in Net conversion but not a practical decrease. So we can say the second part of the hypothesis, “...without significantly reducing the number of students to continue past the free trial”, matches our results because it is not reduced beyond the negative practical significance limit of d_min(-0.0075). Therefor implement the changes.

Be narrower and include only the negative d_min(-0.0075):

This would need a judgment call to implement changes. This may or may not be a practically significant reduced number of students to continue past the free trial and make at least one payment.

Be narrower and less than the negative d_min(-0.0075):

In this case the second part of the hypothesis is proven false. We will significantly reduce the number of students to continue past the free trial and make at least one payment. Therefore do not implement changes.

Follow-Up Experiment

Another possible way to reduce the number of students leaving the free trial early is to do a similar experiment but instead of a message setting ‘clearer expectations’ using a skills test to make sure students meet the prerequisites for a course. If the student does well on the skills test they would continue with the enrollment process. If the student did poorly they would be given the option to take the prerequisite courses or continue to enrollment.

The hypothesis would be if the student passed the skills test they would be better prepared, thus reducing the number of frustrated students who left the free trial—without significantly reducing the number of students to continue past the free trial and eventually complete the course.

The invariant metrics I would choose are Number of cookies, Number of clicks and Click-through-probability. I would not expect these to change from the control to the experimental groups. The evaluation metrics would be Gross conversion, Net conversion and Retention. If the hypothesis is correct I would expect Gross conversion would decrease , Retention to increase and Net conversion to stay the same.

The unit of diversion would be a unique cookie. This cookie would be assigned to new users viewing the course overview page and would be counted once per day. The cookie is a good unit of diversion because it allows the users progression to the next level down the funnel to be calculated as a proportion of the previous step. For example it enables the calculation ot the proportion(Click-through-probability) of users to that view the course overview page that click

the “Start free trial” button. At the time a cookie is created it is randomly assigned to the control or experimental groups.

Sources:

(1) https://docs.google.com/document/d/1ipzp1X8sII_a9UFAhs0d6kYLZQ3Fy4CPlfU1X98u8Lw/edit#
(2) https://classroom.udacity.com/nanodegrees/nd002/parts/00213454013/modules/411033896375460/lessons/4001558669/concepts/39700990110923
(3) http://www.evanmiller.org/ab-testing/sample-size.html
(4) https://classroom.udacity.com/nanodegrees/nd002/parts/00213454013/modules/411033896375460/lessons/4001558669/concepts/39700990160923
(5) https://en.wikipedia.org/wiki/Sanity_check
(6) http://www.statisticssolutions.com/non-parametric-analysis-sign-test/
(7) http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/nonparametrics-tests/understanding-nonparametric-tests/
(8) https://classroom.udacity.com/nanodegrees/nd002/parts/00213454013/modules/411033896375460/lessons/4001558669/concepts/39700990290923
(9) http://www.aaos.org/AAOSNow/2012/Apr/research/research7/?ssopc=1
(10) https://classroom.udacity.com/nanodegrees/nd002/parts/00213454013/modules/411033896375460/lessons/4085798776/concepts/40748587040923
(11)