# Udacity Data Analyst Nanodegree

## Project 7: Design an A/B Test: Udacity Free Trial Screener

## By Stephen Fox
## January 2017

## I. Overview

The objective of this project is to consider the design elements, analyze the results and make recommendations for an A/B test. The specific A/B test under consideration involves a change to the Udacity home page. The home page has two options: “start free trial” and “access course materials”.

Under the current (control) scenario, if the student clicks on “start free trial”, they are prompted to enter their credit card information and they then have access to a 14-day free trial, after which they are automatically charged unless they have cancelled before the trial period ends.

Under the experimental scenario, if the student clicks on “start free trial”, they are asked how many hours per week they can commit to the course. If the answer is 5 hours or more, then they are routed as before (to the credit card screen) for the free 14-day trial. If the answer is less than 5 hours, a message appears stating 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 instead. At that point, the student could then choose to continue with the free trial as before (i.e. be routed to the credit card screen) or they could choose to access the free course materials instead.

The hypothesis was as follows: by making this change, Udacity sets clearer expectations up front and reduces the number of unprepared students (in terms of available time) from taking the free trial, and therefore reduces the number of frustrated students leaving the free trial, without reducing the number who continue past the free trial and ultimately complete the course. Under this scenario, Udacity could improve the overall student experience and improve the coaches’ capacity to support students who are more likely to complete the course.

# II. Experiment Design

## Metric Choice

The unit of diversion in this experiment is a cookie. A list of seven metrics and their respective practical significance levels was provided by Udacity, as follows:

- **Number of cookies:** That is, number of unique cookies to view the course overview page. ($d_{min}$=3000)
- **Number of user-ids:** That is, number of users who enroll in the free trial. ($d_{min}$=50)
- **Number of clicks:** That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). ($d_{min}$=240)
- **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.($d_{min}$=0.01)
- **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. ($d_{min}$= 0.01)
- **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. ($d_{min}$=0.01)
- **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. ($d_{min}$= 0.0075)

The following design decisions were made regarding the utilization of these seven metrics:

**Invariant Metrics**

- **Number of cookies**: this metric should not change between the experiment and control because it measures views of the course overview page, which occurs prior to clicking ‘start free trial’. Therefore, it makes sense to be an invariant and not an evaluation metric.

- **Number of clicks:** this metric should not change between the experiment and control because it measures clicks on the ‘start free trial’ button, which is the step right before the diversion decision. Therefore, it makes sense to be an invariant and not an evaluation metric.

- **Click-through-probability:** this metric should not change between the experiment and control because it measures number of cookies divided by number of clicks. Since both of those metrics should be held constant between the experiment and control groups, as discussed above, it makes sense that their combination in this manner should be constant too. Therefore, it makes sense to be an invariant and not an evaluation metric.

**Evaluation Metrics:**

- **Gross conversion:** According to the experimental hypothesis, one would expect fewer students to enroll in the free trial for the experimental leg of the trial, therefore this metric is expected to be lower for the experimental leg and should be an evaluation metric. Given that it is expected to vary between legs, it is an unsuitable invariant metric. In order to launch the experiment, one would like to see this metric decrease, since that would indicate students who aren't able to commit sufficient time to their studies are being 'weeded out'.

- **Retention:** According to the experimental hypothesis, one would expect fewer students to enroll in the free trial for the experimental leg of the trial, but possibly the number of paying students after 14 days to be unchanged or even increase, and therefore this metric is expected to be different for the experimental leg compared to the control leg and should be an evaluation metric. Given that it is expected to vary between legs, it is an unsuitable invariant metric. In order to launch the experiment, one would like to see this metric increase, since that would indicate that of the fewer students enrolling (due to the 'weeding out' step), the same or more are staying on as paying, satisfied customers after 14 days.

- **Net conversion:** According to the experimental hypothesis, one would expect fewer students to enroll in the free trial for the experimental leg of the trial, but possibly the number of paying students to be unchanged, and therefore this metric is expected to be different for the experimental leg compared to the control leg and should be an evaluation metric. Given that it is expected to vary between legs, it is an unsuitable invariant metric. In order to launch the experiment, one would like to see this metric remain unchanged, since that would indicate that for a given number of students 'clicking start free trial', at least as many remain on as paying, satisfied customers after 14 days.

**Metrics not used:**

- **Number of user-ids:** This metric is not suitable as an invariant metric, since it should be different between the control and experimental groups, since for users that do not enroll, their

user-id is not tracked. It is also not a useful evaluation metric by itself, since it doesn't tell us anything that would let us address whether the experimental hypothesis is true or not.

## Measuring Standard Deviation

Udacity provided baseline data for a single day, as provided in Table 1.

**Table 1: Udacity Baseline Daily Values**

| Data Description | Value |
|---|---|
| Unique cookies to view page per day | 40,000 |
| Unique cookies to click 'start free trial' per day | 3,200 |
| 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 data, standard deviations were analytically estimated for the three evaluation metrics, for a sample size of 5,000 cookies visiting the course overview page, and the results are provided in Table 2.

**Table 2: Analytical Estimates of Standard Deviations for Evaluation Metrics**

| Metric | Probability | N | Standard Deviation |
|---|---|---|---|
| Gross Conversion | 0.20625 | 400 | 0.0202 |
| Retention | 0.53 | 82.5 | 0.0549 |
| Net Conversion | 0.1093125 | 400 | 0.0156 |

For metrics where the unit of analysis (the metric denominator) is not equal to the unit of diversion (cookies in this experiment), then the actual variability can be several times higher than the analytical estimate. Therefore, in the case of retention, where the unit of analysis is user-ids, the analytic estimate of variability is probably lower than the empirical variability and thus an empirical estimate might be worthwhile, if there is sufficient time to determine one. For the other two metrics (gross conversion and net conversion), the unit of analysis is cookie and therefore the analytical estimate given in Table 2 should be reasonably accurate.

## Sizing

### Number of Samples vs. Power

An online tool[1] was used to estimate the page views required to adequately power the experiment, assuming an alpha of 0.05 and a beta of 0.2. The sample size per variation value from the website was converted into page views by using the traffic assumptions provided in Table 1 by Udacity. The resulting page view requirements for each evaluation metric are given in Table 3.

**Table 3: Number of Page Views Required to Provide Sufficient Statistical Power**

| Evaluation Metric | Page Views Required* |
|---|---|
| Gross Conversion | 645,875 |
| Retention | 4,741,212 |
| Net Conversion | 685,325 |

*Example calc.: Retention: 39,115 sample per variation x 2 variations x 40,000 page views / 660 enrollments

### Duration vs. Exposure

The evaluation metric with the highest page view requirement, retention, should be used for design purposes. At 4,741,212 page views required, the experiment would take 119 days[2] assuming 100% of the Udacity traffic was routed to the experiment. Even if one were willing to route all the traffic to the experiment, 119 days (17 weeks) far exceeds the requirement that the experiment only run 'a few weeks'. Therefore, retention will be dropped as an evaluation metric.

The next limiting metric is net conversion, requiring 685,325 page views. At 100% of traffic, it would take 18 days to complete the experiment. I think the change is risky enough that you wouldn't want to run all the traffic to the experiment. It is risky because Udacity will actively attempt to discourage certain students from signing up for a free trial, which could ultimately impact the company's revenues. Since 18 days is only 2.5 weeks, I believe the percent routed to the experiment can and should be reduced from 100%. I suggest routing 50% of the traffic. Doing so increases the amount of time required to 35 days (5 weeks), which I believe still satisfies the requirement that the experiment run no longer than 'a few weeks' while reducing the overall risk that the experiment poses to the business.

[1] http://www.evanmiller.org/ab-testing/sample-size.html
[2] 4,741,212 page views / 40,000 page views per day = 119 days

I believe 50% still feels like a risky amount of traffic to divert, given that revenues are at stake. Ultimately though, it would be a business tradeoff between getting the experiment done relatively quickly versus accepting risk to the existing business.

# III. Experiment Analysis

Udacity provided data showing the results of the experiment, run over 37 days. For each day, experiment and control results were provided for page views, clicks, enrollments and payments. Given the 14-day lag that results for a payment (due to the free trial), only 23 days worth of enrollments and payments data was provided. The daily average and daily standard deviation for these measurements are provided in Table 4, for informational purposes.

**Table 4: Daily Data Statistics for Experiment Results**

| | Control Group | | Experimental Group | |
|---|---|---|---|---|
| **Measurement** | **Daily Average** | **Daily Standard Deviation** | **Daily Average** | **Daily Standard Deviation** |
| Page views | 9,339 | 740 | 9,315 | 708 |
| Clicks | 767 | 68 | 766 | 65 |
| Enrollments | 165 | 30 | 149 | 33 |
| Payments | 88 | 21 | 85 | 23 |

## Sanity Checks

As a sanity check on the experiment, the invariant metrics were calculated and subjected to a hypothesis test to confirm that the control and experimental groups did not differ from each other at the 95% confidence level. For 'number of cookies' and 'number of clicks on "start free trial"', the confidence interval was calculated around the expected fraction of events that are assigned to the control group (i.e. ½ or 0.5). For the 'click-through probability on "start free trial"', the confidence interval is around the expected difference in proportions between the control and experimental groups, which should be zero (0) for an invariant metric. The results of the sanity check is presented in Table 5, where the data indicates that all three invariant metrics pass (i.e. they are not different at the 95% confidence level, when comparing the control and experiment group).

**Table 5: Invariant Metrics Sanity Check Results – 95% Confidence Intervals**

| Invariant Metric | Lower Bound | Upper Bound | Observed | Pass? |
|---|---|---|---|---|
| Number of cookies | 0.4988 | 0.5012 | 0.5006 | Yes |
| Number of clicks on 'start free trial' | 0.4959 | 0.5041 | 0.5005 | Yes |
| CTP* on 'start free trial' | -0.0013 | 0.0013 | 0.0001 | Yes |

*CTP = click-through probability

## Result Analysis

### Effect Size Tests

For the two evaluation metrics under consideration, a 95% confidence interval around the difference between the experiment and control groups was calculated. The results are given in Table 6. Since the confidence interval on 'gross conversion' did not include zero, it is statistically different for the control group and experiment group. On the other hand, the confidence interval on 'net conversion' does include zero, and therefore it is not statistically different between the two groups at the 95% confidence level. Practical significance exists when the confidence interval does not include the practical significance boundary. The practical significance for gross conversion was given as 0.01, and since the absolute value on the upper bound of the 'gross conversion' confidence interval exceeds this amount (i.e. 0.012 > 0.01), the 'gross conversion' is also of practical significance. By virtue of not being statistically significant, the 'net conversion' is not of practical significance.

**Table 6: 95% Confidence Intervals on Evaluation Metrics**

| Evaluation Metric | Lower Bound | Upper Bound | Statistical Significance? | Practical Significance |
|---|---|---|---|---|
| Gross Conversion | -0.0291 | -0.0120 | Yes | Yes |
| Net Conversion | -0.0116 | 0.0019 | No | No |

These results are consistent with what was surmised about these metrics in the experiment design section, namely that gross conversion should be lower in the experiment group, due to the 'weeding out' effect. The fact that both the upper and lower bounds of the confidence interval are negative indicates that this is the case. It was surmised that 'net conversion' should stay unchanged, and given that there is no statistical difference between the experiment and control groups, it appears to be unchanged, at least at the 95% confidence level.

### Sign Tests

A sign test was conducted using an online tool[3], to confirm the results of the confidence interval testing presented in Table 6. The results of the sign test are given in Table 7.

[3] http://graphpad.com/quickcalcs/binomial1/

**Table 7: Sign Test for Statistical Significance of Evaluation Metrics**

| Evaluation Metric | p-value | Statistical Significance? |
|---|---|---|
| Gross Conversion | 0.0026 | Yes |
| Net Conversion | 0.6776 | No |

Since the p-value for 'gross conversion' (0.0026) is less than alpha (0.05), it is statistically significant. On the other hand, since the p-value for 'net conversion' (0.6776) is more than alpha (0.05), it is not statistically significant. These results agree with the confidence interval testing presented in Table 6.

**Summary**

In this analysis, I decided not to use the Bonferroni estimate for alpha, because it is known to be very conservative and also because the maximum number of evaluation metrics beings considered here (3 or fewer) was not exceptionally large. For the final analysis, only two evaluation metrics were considered ('gross conversion' and 'net conversion'), since the page view requirement for powering the 3rd candidate evaluation metric ('retention') would have required the experiment to last at least 119 days, which was far in excess of the project scope of 'a few weeks'. Based on the two evaluation metrics selected, a 35-day experiment was required.

The results indicated that the experimental setup, where potential students are prompted to enter their available learning time per week, resulted in a statistically significant reduction in the 'gross conversion' metric. This was consistent with expectations, since fewer students would be expected to enroll in the free trial if prompted to consider the time commitment requirement. The experimental setup did not result in a statistically significant change in the 'net conversion' metric, meaning that the number of students to stay enrolled past 14 days appears to be unchanged between the two groups, which is a positive business outcome.

## Recommendation

I believe the results warrant implementing a gradual ramp up of the experimental design, although ultimately it is a judgment call for a leader with better insight into the cost and current utilization of the coaching staff, for the reasons now described:

The fact that 'net conversion' is not statistically different suggests that the number of revenue generating customers will not decrease by making this change. On the other hand, the fact that 'gross conversion' is statistically lower under the new design suggests that coaching resources

will be freed up to assist those students that are most likely to continue through the program. This could result in the total Udacity offering improving for those students most likely to complete courses, assuming coaching resources are generally operating at full capacity currently. However, if coaching resources are currently not operating at full capacity, freeing them up further is unlikely to result in any worthwhile improvements to the business's performance. If this is the case, then making this change and exposing the business to some risk (albeit minimal) might not be worth it.

# IV. Follow-Up Experiment

## Design Overview

In an attempt to reduce the number of frustrated students who cancel early in the course, I recommend an experiment centered on enhancing the forum experience for students. Based on my personal experience with Udacity (several courses, two completed Nanodegrees, and a 3rd Nanodegree (this one) in progress), it took me many months to really figure out the usefulness of the forums, and I think the following experiment would help greatly increase the learning rate for students regarding the value of the forums.

The experiment would work as follows: once customers have initiated a free trial and 'checked out' under the current system architecture, the control group would proceed with the course as usual whereas the experimental group would be provided a very brief (1-2 slides) overview of using the forums and the types of questions one can ask there, highlighting the generally very quick response time to posted questions. After this brief overview, the experimental group would be routed to the same starting point as the control group (i.e. the 1st slide in the course or Nanodegree).

## Hypothesis

The hypothesis of this follow-up experiment is as follows: the primer on the forums will result in the experimental groups being more prolific forum users, which in turn will improve their performance on the courses, reduce their frustration and enhance retention.

## Metric Choices

I would recommend user-id as the unit of diversion, for the reason that the experiment is limited to those students who have initiated the free trial, and hence they should all be logged in during the experiment (this is in contrast to the previous experiment, where the process initiated at the course overview page, prior to logging in).

The metrics I would measure are centered around trying to measure whether there are differences in how users utilize their time on the Udacity site. Do the two groups view the forums more or less frequently? Do the two groups proceed at a different pace through the course material? Finally, is there a difference between the two groups in retention?

At a minimum, I would recommend measuring the following metrics:

- **Number of user-ids:** number of user-ids to complete checkout
- **Number of forum clicks:** number of times a user clicks links in the forums
- **Number of unique forum users:** number of unique user-ids to click on at least one forum link
- **Number of forum posts:** number of times a user posts in the forums
- **Number of unique forum posters:** number of unique user-ids to post at least once in the forums
- **Number of classroom slides click throughs:** number of times a user clicks to the next classroom slide (an imperfect measure of progress through the classroom material)
- **Retention:** 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.

I am by no means advanced at computer system architecture, so it is unclear to me how easy or difficult it is to measure how much time users are spending on various slides or in the forums, but based on feedback from the engineering team regarding feasibility, I think it would be very interesting and insightful to track metrics on how the users are spending their time, to see if there is a difference between the two groups (e.g. can we measure the experimental group spending more time on the forums and / or less time on each classroom slide?).

The ultimate evaluation metric would be ‘retention’, where one would expect to see an increase in this metric for the experimental group, were the hypothesis to hold true.