P7: A/B Testing - Udacity Free Trial Screener

Student: Andy Miller Udacity's Data Analyst Nanodegree

Please find below my qualitative write-up with the inclusion of the quantitative detail.

Experiment Design

Metric Choice

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

(Invariant Metric Choice | Evaluation Metric Choice)

1. **Number of Cookies** *(No. of unique cookies to visit the course overview page)*: As our unit of diversion here is cookies, it will be evenly distributed between the control and experiment groups. I will be testing to ensure this metric doesn't vary too significantly between the control and experiment groups.

2. Number of User IDs *(No. of users who enroll in the free trial)*: This metric is was not a good invariant metric choice, as it is the subject of what is being influenced here, that is, signing up students in the free trial. Further, I did not select this as an evaluation metric as I believe that the metric would be subjected to too much variation daily that could not be easily understood. There were other choices for evaluation metrics available that I believe were better suited.

3. **Number of Clicks** *(No. of unique cookies to click the "Start Free Trial" button)*: Similar to metric #1 above, this should be evenly distributed between the groups and provides a good metric to monitor and check. Similarly, I will test to ensure this metric doesn't vary too significantly between the control and experiment groups

4. Click-Through-Probability *(No. of unique cookies to click the "Start Free Trial" button / No. of unique cookies to visit the course overview page)*: As this is the ratio of the other two invariant metrics chosen, this metric would provide us with a sanity check on the data to ensure even distribution. However, this metric is a direct analog of the other two metrics already chosen. I originally had chosen this as an invariant metric, but in the course of my testing, I decided to stick with the other two metrics, as for me, it would be easier to review and understand the results of my sanity checks.

5. **Gross Conversion** *(No. of user IDs to complete checkout and enroll in trial / No. of unique cookies to click "Start Free Trial" button)*: This metric is suitable, as it is dependent upon the experiment and allows us to determine if there was a change to one of the key factors being discussed in the hypothesis, that is enrollments. As the hypothesis is seeking to set expectations with users (re: time commitments) and potentially reduce the number of users that enrolled, and theoretically allow Udacity to improve support capacity and overall experience, while reducing frustrated users prior to frustration.

   I expect to see a decrease in the experiments metric, as compared with the control group, as there will be fewer students that complete checkout and enrollment, but the same number of users clicking the "Start Free Trial" button.

6. **Retention*** *(No. of User IDs to remain enrolled past 14-day boundary (and pay) / No. of user-ids to complete checkout)*: As mentioned above, users in the experiment group become aware of the time commitment up-front. Retention, as an evaluation metric, takes two significant measures into account: enrollment to a point where the user pays, and users that complete upfront checkout for trial.

   For this experiment, I expect to see a significant increase in the experiment group's retention rate, as fewer users will enroll, but these users will be more committed, meaning they will stay past the 14-day boundary. * *Noted in section on "Sizing" this evaluation metric was removed from consideration. Refer there for more information.*

7. **Net Conversion** *(No. of User IDs to remain enrolled past 14-day boundary (and pay) / No. of unique cookies to click "Start Free Trial" button)*: Finally, net conversion is a combination of Gross Conversion and Retention, whereby we have the payment component over the larger number of how many actually clicked the "Start Free Trial" button.

   I expect this number to be significantly smaller than the other two evaluation metrics, as the denominator (click button) is much larger than Retention's denominator and the numerator (payment) is much smaller than Gross Conversion's numerator (enrollment). For this experiment, I would hope this metric increases or stays the same.

Launch Criteria

In order to launch this experiment, I must see a statistically and practically significant change with gross conversion and retention metrics, where gross conversion is a negative change and retention is a positive change. With regards to net conversion, a statistically significant change is not critical, but the metric must remain the same or have a positive increase.

Measuring Standard Deviation

Gross Conversion: SQRT(0.20625*(1-0.20625)/(5000*(3200/40000))) = **0.02023**
Retention: SQRT(0.53*(1-0.53)/(5000*(660/40000))) = **0.0560**
Net Conversion: SQRT(0.1093125*(1-0.1093125)/(5000*(3200/40000))) = **0.0156**

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

For both Gross Conversion and Net Conversion, the "Unit of Analysis" is the same as the "Unit of Diversion", indicating that the analytical estimate will be similar to the empirical variability. For Retention, we know that the denominator is user IDs that have enrolled (checked-out), which is not the same as the "Unit of Diversion", indicating that the analytical and empirical estimates would be different.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of page views you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Page views" quiz.)

I did not utilize Bonferroni's correction during my analysis. I believe my evaluation metrics are correlated and I believe that Bonferroni would be too conservative here. Further, in this case, I have decided that I want ALL of my metrics to be statistically significant in order to make a decision, therefore Bonferroni is not needed.

Regarding the sizing of this test, I utilized Evan Miller's "Sample Size Calculator" to determine the number of page views I would need for each evaluation metric. Results are below:

Gross Conversion

α: 5% β = 20% 1−β: 80%
dMin (Practical Significance) = 1%
Conversion Rate: 0.20625
Samples needed: 25,835
Page views needed: 2* (25,835 / (3200/40000)) = **645,875**

Retention

α: 5% β = 20% 1−β: 80%
dMin (Practical Significance) = 1%
Conversion Rate: 0.53
Samples needed: 39,115
Page views needed: 2* (39,115 / (660/40000)) = **4,741,200**

NOTE: Upon seeing the necessary page views to maintain the statistical power for the Retention metric, I quickly realized that the 40K daily page views was going to make this a very long experiment, and in my eyes, very impractical. If devoting 100% of traffic to the experiment, the experiment would last 118 days. This is much too long and presents more problems (especially introducing variability that we might not be able to explain or understand), therefore I eliminated this as an evaluation metric.

Net Conversion

α: 5% β = 20% 1−β: 80%
dMin (Practical Significance) = .75%
Conversion Rate: 0.1093125
Samples needed: 27,413
Pageviews needed: 2* (27,413 / (3200/40000)) = **685,325**

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. (These should be the answers from the "Choosing Duration and Exposure" quiz.) Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

I determined that I would divert 70% of traffic. I considered many factors when making this decision:

1. Safety and Risk – Changing the way we push students into a formal, paid trial is a significant change in the formula for onboarding legitimate students (that pay for support), therefore I did not want to move 100% of traffic as it would be important to see how users react and the process performs technically.

Additionally, I wanted to hold-back some portion of data to compare against, specifically making sure that changes are due to any seasonality or time of day issues (as two simple examples).

2. Time – I wanted to keep the experiment open long enough to have good tracking data, but also short enough that erroneous variability that cannot be easily understood or explained isn't introduced, as well as, ensure we had results within a reasonable amount of time. Through trial and error, I decided that 25 days is reasonable to capture data for these two evaluation metrics and meets my requirement of keeping the experiment to less than 30 days.

Number of Days Needed to Run Testing
Gross Conversion: 645,875 /40,000 * (1/0.7) = 23.067 Days
Net Conversion: 685,325 /40,000 * (1/0.7) = **24.4759 Days**

Overall, this experiment is not affecting current users, content or Udacity's current revenue, therefore I do see this as a high-risk experiment, however, as noted above, I chose not to divert 100% of traffic to manage the risk that is present.

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.) For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. Do not proceed to the rest of the analysis unless all sanity checks pass.

Number of Cookies

P(Cookie in Control or Experiment Group) = 0.5
SE: SQRT(0.5*(1-.5)/(345543+344660)) = 0.0006
Margin of Error (m) = 1.96 * 0. 0006 = 0.0012
CI - Upper: 0.5 + 0. 0012 = **0.5012**
CI - Lower: 0.5 - 0. 0012 = **0.4988**
Observed Value (Numerator is control): 345543/(344660 + 345543) = **0.5006**
PASSES CHECK - Fits within Confidence Interval

Number of Clicks

P(Cookie in Control or Experiment Group) = 0.5
SE: SQRT(0.5*(1-.5)/(28378+28325)) = 0.0021
Margin of Error (m) = 1.96 * 0. 0021 = 0.0041
CI - Upper: 0.5 + 0. 0041= **0.5041**
CI - Lower: 0.5 - 0. 0041= **0.4959**
Observed Value (Numerator is control): 28378/(28325+28378) = **0.5005**
PASSES CHECK - Fits within Confidence Interval

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)

Summary Data Table

	Control	**Experiment**	**TOTAL**
Clicks	17293	17260	34553
Enrollments	3785	3423	7208
Payments	2033	1945	3978

Gross Conversion:

Pooled Probability: 7208 /34553 = 0.2086
Pooled SE: SQRT(0.2086 *(1-0.2086)*((1/17293)+(1/17260))) = 0.0044
Margin of Error (m): 1.96*0. 0044=0.0086
Dhat: (3423/17260) - (3785/17293) = -0.0206
CI - Upper: -0.0120
CI - Lower: -0.0291
Statistically Significant: Yes – CI does not contain zero
Practically Significant: Yes – CI does not contain Dmin value of 0.01

Net Conversion:

Pooled Probability: 3978 / 34553 = 0.1151
Pooled SE: SQRT (0.1151 *(1-0.1151)*((1/17293)+(1/17260))) = 0.0034
Margin of Error (m): 1.96*0. 0034= 0.0067
Dhat: (1945/17260)-(2033/17293) = -0.0049
CI - Upper: 0.0018
CI - Lower: -0.0116
Statistically Significant: No – CI does contain zero
Practically Significant: No – CI does not contain Dmin value of +/-0.0075

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

Gross Conversion

- Number of Successes: 4
- Number of Trials: 23
- P = .5
- Two-tail P value is **0.0026 which IS statistically significant (less than alpha).**

Net Conversion

- Number of Successes: 10
- Number of Trials: 23
- P = .5
- Two-tail P value is **0.6776 which IS NOT statistically significant (greater than alpha).**

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

As mentioned above, I did not utilize Bonferroni's correction during my analysis. I believe my evaluation metrics are correlated and I believe that Bonferroni would be too conservative here. Further, in this case, I have decided that I want ALL of my metrics to be statistically significant in order to make a decision, therefore Bonferroni is not needed. I do understand the importance of insuring that the launching of multiple metrics can result in a false positive (especially as the number of metrics launched increases). Though, I would only launch if all evaluation metrics show a statistically significant change, above my practical significance level (to ensure the return is significant enough to justify the investment). Correction would be applied if I were using OR on all metrics, but not when testing for AND of all metrics (reference).

Recommendation

Make a recommendation and briefly describe your reasoning.

My recommendation is to NOT launch / implement this new adjustment to the process for onboarding students. My recommendation centers around the results of the evaluation metrics:

Gross Conversion

Upon review of the experiment and results, we did confirm that this metric did in fact get reduced, which was expected as we would likely be turning away more users. We know, based upon the hypothesis that this is a good result, as support costs would be reduced and quality likely improved, thus improving the student experience.

Net Conversion

Unfortunately, with this metric, we did not have the same result of significance. It was our hope and expectations that there would be a positive change here (i.e., number of actual payments was increased), but this did not result. We actually saw negative confidence intervals, and never met our very small practical significance leave of 0.0075. Overall, the users from the experiment must not have been influenced by the warning (which I expected to create a "challenge" for the student).

Based on this analysis, we have a positive and negative result (i.e., Net Conversion where there was no statistical significance) with our evaluation metrics. When reviewing the initial hypothesis, this experiment has not increased the number of paying users, therefore making this feature un-launchable.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

I have plenty of experience with participating in software / learning trials, which all come in various sizes and features. For example, some trials are 14-days, 30-days or 90-days (I have see 14 and 30 as the most common). Additionally, I see products that are enabled during the trial period in a limited fashion (i.e., features reduced / disabled), as well as fully-featured trials. With regards to Udacity, I would like to test the length of the trial. For me, this was not a major barrier, because I tend to be committed and enthusiastic about learning, but I know this to be an issue with others. I believe it would be interesting to see whether a 30 or even 90-day trial would effect the number of students that stayed longer-term (over the "hump period" of 45 days). I believe that users get excited and engaged for just the first 1-3 days, then they get busy for a week and realize their trial is almost over, causing them to just concede.

Thus, the experiment is to provide a group of experimental users with a 30-day trial. My hypothesis is that this would actually cause greater adoption of the product and thus more revenue for the user long term. My unit of diversion would be cookies, as I need to advertise to a new user before they have enrolled. I would want to track number of cookies that view the trial overview page as an invariant metric, since the unit of diversion is cookie, and because this would be prior to the user being aware of the specifics of the trial. My evaluation metric would be “Long-Term Conversion Rate” which would be: No. of customers staying at least 45 days (meaning at least one payment) / number of users enrolling in the trial. One significant point about my evaluation metric is the “45 days”. I determined that this was the time needed to get a user “over the hump”. Long enough past the 30-day trial where they have already realized they are paying for the education, but short enough to not draw this experiment out too long.

If Long-Term Conversion Rate increases in a statistically and practically significant manner by the conclusion of the experiment, I would launch this across all of Udacity, as it’s a fundamental change and would drive more students to stick with it (improving their life with better education) and improving the “bottom line” of Udacity as they retain customers that just needed a few more weeks to get over that initial hump.