# Experiment Design

## Metric Choice

Invariant Metrics:

* **Number of cookies:** number of unique cookies to view the course overview page. (dmin=3000)

The number of cookies should be kept the same between experiment and control group

* **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). (dmin=240)

Since, the change is after user clicks the ‘Start free trial’, the number of click will not be affected by the new feature. Hence, the number of clicks should be invariant metric.

* **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. (dmin=0.01)

Same as the number of click, the click-through-probability will not be affected by the change. Hence, should be the same between experiment and control group.

Evaluation Metrics:

* **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. (dmin= 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. (dmin=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. (dmin= 0.0075)

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.

Therefore, the gross conversion may reduce. Retention is expected to increase, and net conversion should be about the same. Hence, these three should be evaluation metrics.

## Measuring Standard Deviation

* **Gross conversion:**

N = 5000\* Click-through-probability on "Start free trial" = 5000\*0.08 = 400

SE = ((Probability of enrolling, given click)\*(1- Probability of enrolling, given click)/N)0.5

= (0.20625\*(1-0.20625)/400)0.5

= 0.02023

* **Retention:**

N = 5000\* Enrollments per day/ Unique cookies to view course overview page per day

= 5000\*660/40000

= 82.5

SE = ((Probability of payment, given enroll)\*(1- Probability of payment, given enroll)/N)0.5

= (0.53\*(1-0.53)/82.5)0.5

= 0.05495

* **Net conversion:**

N = 5000\* Click-through-probability on "Start free trial" = 5000\*0.08 = 400

SE = ((Probability of payment, given click)\*(1- Probability of payment, given click)/N)0.5

= (0.1093125\*(1-0.1093125)/400)0.5

= 0.01560

For all the evaluation metrics, the analytic estimate would be lower than the empirical variability. Since, the unit of division (cookie) is different from the dominator of the evaluation metric.

## Sizing

### Number of Samples vs. Power

Bonferroni is not applied, because these metrics are closely related to each other, and Bonferroni would be too conservative.

* **Gross conversion:**
  + Baseline conversion rate = 0.20625
  + Minimum detectable effect = 0.01
  + 1 – β = 0.8
  + α = 0.05
  + Sample size = 25835\*2
  + # of page view = 25835\*2/0.08 = 645875
* **Retention:**
  + Baseline conversion rate = 0.53
  + Minimum detectable effect = 0.01
  + 1 – β = 0.8
  + α = 0.05
  + Sample size = 39115\*2
  + # of page view =39115\*2/(660/40000) = 4741212
* **Net conversion:**
  + Baseline conversion rate = 0.109313
  + Minimum detectable effect = 0.0075
  + 1 – β = 0.8
  + α = 0.05
  + Sample size = 27413\*2
  + # of page view = 27413\*2/0.08 = 685325

Pageviews required is maximum of pageviews required for Gross Conversion, Retention, Net Conversion. Therefore, the required pageviews is 4741212.

### Duration vs. Exposure

If the fraction of traffic exposed is 1, the length of experiment is 119 days in order to collect 4741212 pageviews. This experiment duration is too long. A way to reduce the experiment length is getting rid of the measure metric – retention. After the change, the number of pageviews and duration decrease to 685325 and 18 days, respectively. If we change the fraction of traffic exposed to 0.5, the length of experiment would increase to 35 days.

# Experiment Analysis

## Sanity Checks

* **Number of clicks - passes**
  + Expected value = 0.5
  + Observed value = 28378/(28378+28325) = 0.500467347
  + sd = SQRT(0.5\*(1-0.5)/(28378+28325)) = 0.002099747
  + m = 1.96\*sd = 0.004115504
  + CI\_LB = sd – m = 0.495884496
  + CI\_UP = sd + m= 0.504115504
* **Number of cookies - passes**
  + Expected value = 0.5
  + Observed value = 345543/(345543+344660) = 0.500639667
  + sd = SQRT(0.5\*(1-0.5)/(345543+344660)) = 0.000601841
  + m = 1.96\*sd = 0.001179608
  + CI\_LB = sd – m = 0.498820392
  + CI\_UP = sd + m= 0.501179608
* **Click-through-Probability - passes**
  + Expected value = 28378/345543 = 0.82125814
  + Observed value = 28325/344660 = 0.82182441
  + sd = SQRT(0.82125814\*(1-0.82125814)/345543) = 0.000467068
  + m = 1.96\*sd = 0.000915454
  + CI\_LB = sd – m = 0.08121036
  + CI\_UP = sd + m= 0.083041267

All the sanity checks pass, since the observed values fall in the confidence intervals.

## Result Analysis

### Effect Size Tests

A 95% confidence interval around the difference between the experiment and control groups are shown in table below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **dmin** | **dobs** | **CI\_LB** | **CI\_UB** | **Satistically Significant** | **Practically**  **Significant** |
| Gross Conversion | 0.01 | -0.0205 | -0.0291 | -0.0120 | Yes | Yes |
| Net Conversion | 0.0075 | -0.048 | -0.0116 | 0.0019 | No | No |

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

* Gross Conversion
  + Number of successes = 4
  + Number of experiment = 23
  + P-value = 0.0026 < 0.05
  + Statistical significant: Yes
* Net Conversion
  + Number of successes = 10
  + Number of experiment = 23
  + P-value = 0.6776 > 0.05
  + Statistical significant: No

### Summary

In this 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. 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. The unit of diversion is a cookie.

Three invariant metrics: number of Cookies, number of clicks on "start free trial", and click-Through-Probability, were chosen for validation and sanity checking. While gross conversion retention and net conversion are served as evaluation metrics. Under the null hypotheses, the gross conversion will reduce. Retention is expected to increase, and net conversion should be about the same. Because testing retention would prolong the experiment to 3 months long. We decided to only consider the changes in gross and net conversion.

Bonferroni correction is not applied in this case. Since, these two metrics are highly correlated. And Bonferroni would be too conservative in this case.

Both the effect size hypothesis tests and the sign tests are carried out for the two evaluation metrics: gross and net conversion. 95% CI was used. Both tests yield to the same results. We found out that for the gross conversion, the difference in control and experiment groups is both statistically and practically significant. For net conversion, the difference is nether statistically or practically significant. These results is consisted with our expectations: gross conversion would drop, while net conversion remain the same.

## Recommendation

The results obtained from the effect size hypothesis tests show the difference in gross conversion is significant and the change in net conversion is insignificant. These are consistent with our expectation under the hull hypothesis. Therefore, this change could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course. And we suggest to launch this change.