# 1 Experiment Design

## 1.1 Metric Choice

*Question 1: List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)*

Invariant metrics: number of cookies, number of clicks

Evaluation metrics: gross conversion, retention, net conversion

*Question 2: 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.*

* Number of cookies: That is, number of unique cookies to view the course overview page. (dmin=3000).As it is the unit of diversion and the cookies were randomly assigned to the experiment group and control groups, there should be roughly the same number of cookies in each group. It is a population sizing invariant.
* Number of user-ids: That is, number of users who enroll in the free trial. (dmin=50): It can be a population sizing invariant if the users don’t change their cookies too often, as there will be the same number of user-ids in experiment group and control groups . As the users’ browsing behaviors are not guaranteed, number of user-ids won’t be chosen in this analysis.
* 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 experiment was conducted after the users clicked the button, the users’ number of clicks should not be affected. As a result, it is an invariant matrix.
* 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). As the number of clicks and the number of unique cookies are invariant, the click-through-probability is also an invariant. However, it will not be chosen in this analysis as the relevant information is already covered in the number of clicks and the number of unique cookies separately.
* 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). The matrix evaluates if the questions asked in the experiment group changed the users’ free-trial enrollment behaviors, which can be used as an evaluation matrix. As the questions asked are supposed to discourage less committed students from enrolling in the free-trial, it is expected that the experiment group has a lower gross conversion.
* 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). This is an evaluation matrix. It evaluates if the free-trial users will continue as paid users. As the questions asked are supposed to discourage less-committed students, the enrolled students in the experiment group are more committed and should have a higher rate of retention.
* 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). This is an evaluation matrix and it evaluates directly if the questions asked would help increase the retention rate of all the unique cookies. As the questions asked are supposed to discourage less-committed students to enroll, the students in the experiment group are expected to be more committed, more likely to enroll in the free-trial and continue to pay, so they are supposed to have a higher net conversion. It is a combination of gross conversion and retention.

All the three evaluation matrices should be measured. Although net conversion is a direct evaluation of the experiment’s final results, as it is a combination of gross conversion and retention, which are supposedly moving opposite to each other. As a result, it is also worth measuring gross conversion and retention to find out the causes of the change of net conversion.

## 1.2 Measuring Standard Deviation

*Question 1: List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz)*

Calculation: sqrs(p(1-p)/n) \* sqrt(40000/5000)

(p: the probability, n: the total sample size with a 40,000 pageviews per day)

(converted to *5000 pageviews as specified in the quiz*)

Gross conversion: 0.02023  
Retention: 0.05495  
Net conversion: 0.015601

*Question 2: 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 gross and net conversion, as the unit of analysis and the unit of diversion are number of unique cookies clicking the button, the variability tend to be closer to the analytical estimate. For retention, however, as the unit of analysis and the unit of diversions are different, the variability will not be close to analytical estimate, so empirical estimate might be needed.

## 1.3 Sizing

### 1.3.1 Number of Samples vs. Power

*Question 1: Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)*

I will not use Bonferroni correction. As the three matrices are highly correlated to each other, using Bonferroni correction will be too conservative.

The number of pageviews needed for the three matrices are shown in the table below. The sample size needed were calculated based on the base conversion rate and d\_min, using an online calculator[[1]](#footnote-1).

It was noticed that if retention is used as an evaluation matrix, even if all the traffic were diverted to the experiment, the experiment would still take 4 months, which is very unrealistic. As a result, retention will not be used as the evaluation matrix. The number of pageviews needed to power the experiment is 685,325.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Gross conversion** | **Retention** | **Net conversion** |
| Base conversion rate | 0.20625 | 0.53 | 0.1093125 |
| d\_min | 0.01 | 0.01 | 0.0075 |
| Sample size needed for one group | 25,835 unique cookies | 39,115 free\_trial enrollments | 27,413 unique cookies |
| **Unique pageviews need for the control and experiment groups** | **25,835/0.08\*2**  **= 645,875** | **39,115/660\*40000\*2**  **= 4,747,212** | **27,413/0.08 \*2**  **= 685,325** |
| Days needed if full traffic is diverted to the experiment | 645,875/40,000  = 17 | 4,747,212/40,000  =  119 | 685,325/40,000  = 18 |

### 1.3.2 Duration vs. Exposure

*Question 1: 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.)*

I would divert 80% of the traffic to the experiment, which is 40,000 \* 80% = 32,000 unique cookies to view page per day. 22 days are needed to achieve the 685,325 pageviews required, which is reasonable.

*Question 2: Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?*

The experiment has some risk as it affects the behaviors of new users, but the risk is not very high as it does not affect currently enrolled users. As a result, a higher traffic diversion rate (but not 100%) is chosen for the experiment. It should also be noted that if this is the first time that a similar experiment has been conducted.

# 2 Experiment Analysis

## 2.1 Sanity checks

*Question 1: 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.*

The table below shows that the observed value of number of cookies and number of clicks are within the confidence interval at the 95% confidence level, so both of them passed the sanity check.

|  |  |  |
| --- | --- | --- |
|  | Number of cookies | Number of clicks |
| Control group | 345,543 | 28,378 |
| Experiment group | 344,660 | 28,325 |
| Total | 690,203 | 56,703 |
| P(pool) | 0.5 | 0.5 |
| SE | 0.0006018 | 0.0021 |
| Margin of error = SE\*1.96 | 0.0011796 | 0.0041 |
| Confidence Interval =  (p – margin of error, p + margin of error) | [0.4988, 0.5012] | [0.4959, 0.5041] |
| Observed value  Control group/ Total | 0.5006 | 0.50046 |
| Result | Passed the sanity check | Passed the sanity check |

## 2.2 Result Analysis

### 2.2.1 Effect Size Tests

*Question1: 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.*

The table blow shows that Gross Conversion’s confidence interval is statistically and practically significant, while the Net Conversion’s confidence interval is neither statistically nor practically significant.

|  |  |  |
| --- | --- | --- |
|  | Control Group | Experiment Group |
| Clicks | 17,293 | 17,260 |
| Enrolment (free-trial) | 3,785 | 3,423 |
| Payment | 2,033 | 1,945 |
| Gross Conversion | 0.218875 | 0.198320 |
| Net Conversion | 0.117562 | 0.112366 |

|  |  |  |
| --- | --- | --- |
|  | Gross Conversion | Net Conversion |
| d\_min | 0.01 | 0.0075 |
| Pooled probability | 0.2189 | 0.1151 |
| SE = sqrt(p\*(1-p)\*(1/click\_control +1/click\_exp) | 0.004371 | 0.003434 |
| Margin of error = SE\*1.96 | 0.008568 | 0.0067 |
| d = experiment - control | -0.02055 | -0.0049 |
| Confidence Interval =  [d – margin of error, d + margin of error] | [-0.0292, -0.0120] | [-0.0115, 0.0019] |
| Statistical significance (CI does not include 0) | Significant | Not significant |
| Practical significance (CI does not include d\_min) | Significant | Not significant |

### 2.2.2 Sign Tests

*Question 1: 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.*

The table below show that the hypotheses that gross conversion is statistically and practically significant while the net conversion is neither statistically nor practically significant were accepted.

|  |  |  |
| --- | --- | --- |
|  | Gross conversion | Net conversion |
| Number of success | 4 | 10 |
| Number of trials  (from 11 October to 2 November) | 23 | 23 |
| Probability | 0.5 | 0.5 |
| Two-tailed p-value | 0.0026 < alpha level 0.025 | 0.6776 > alpha level 0.025 |

### 2.2.3 Summary

*Question 1: 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.*

I did not use Bonferroni correction, as the evaluation matrices are highly correlated, the Bonferroni correction will be too conservative. Besides, in this experiment, we want both matrices to be significant, while Bonferroni’s purpose is to control false positive.

The results of effect size tests and sign tests are in line with each other. The gross conversion is both statistically and practically significant while the net conversion is neither statistically or practically significant. This means that adding the questions significantly reduced the gross conversion but did not significantly affect net conversion.

## 2.4 Recommendation

*Question 1: Make a recommendation and briefly describe your reasoning.*

The change should not be launched. The change did reduce the gross conversion as expected, which means the change discouraged less committed students from enrolling in the free trial. This can reduce the cost of Udacity as less resources can be put on these less committed students who are less likely to generate profits in the futures, that is, the students who would have signed up for the free trial but not become paid students.

However, the net conversion is not affected, which means the change did not bring more profits. One possibility is that, although the “committed” students who saw the reminder and still decided to sign up for the free trial are more likely to continue to the paid session, the increase of retention is not enough to compensate for the loss of student registration in the first place.

# 3 Follow-up experiment

*Question1: 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.*

In the previous experiment, the change discouraged the less-committed students to enroll in the free-trial, so the resources on these students who are less likely to bring profit in the future are saved. However, the change did not bring more profits as the net conversion did not increase significantly. In the following-up experiment, I would like to directly intervene the tree-trial period to increase the students’ retention rate, so that the net conversion can also be increased.

My hypothesis is that by providing more easily accessible studying tips for some of the difficult problem sets/projects in the free-trial period, the students will gain a more positive learning experience, so they are more likely to decide to process to the paid period and the retention rate will increase. If the total number of cookies that clicks the button is invariant, the net conversion should also increase.

Unit of diversion: as the experiment is conducted to see if the proposed change has influence on the students who have already enrolled in the free-trial period with and each of these students has a unique user-ID, the user-ID is the unit of diversion.

Invariant metrics: user-ID is invariant metric, as they are randomly assigned to control group and experiment group, each group with roughly the same number of IDs. Number of cookies is invariant if each cookie corresponds to a unique user-id. If the number of cookies is invariant, the number of clicks should also be invariant, as the experiment is conducted after the clicks and the number of clicks is not affected.

Evaluation metric: retention rate and net conversion are the evaluation metrics as stated in the hypotheses.

References:

<http://www.evanmiller.org/ab-testing/sample-size.html>

<https://en.wikipedia.org/wiki/Sign_test>

<https://graphpad.com/quickcalcs/binomial1.cfm>

1. http://www.evanmiller.org/ab-testing/sample-size.html [↑](#footnote-ref-1)