

## A/B Testing Experiment

Udacity has processed an experiment concerning 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.

Initial Unit of Diversion: Cookie

## Experiment Design

### Metric Choice

#### Invariant Metrics

##### Number of Cookies:

It is the number of unique cookies to visit the course overview page. Since the Unit of Diversion is cookies, it is very well suited for Invariant Metrics and it is evenly distributed across the control and experiment group (random). Considering the latter, it cannot be used as an Evaluation metrics.

##### Number of Clicks:

It is the number of unique cookies to click the "start free trial" button. Since, the page is asking for the number of hours the student can devote to the course, which appeared after clicking the "start free trial" button, the course overview page remains the same for both the control and experiment group. Thus, there is an equal probability of clicking the "start free trial" button by both the control and experiment group and it can be used for sanity checking as an Invariant metric. Since the experiment popup appears after clicking the "start free trial" button, it cannot be used as an evaluation metrics as it wouldn't notice any change between the control and experiment group.

#### Evaluation Metrics

##### Gross Conversion:

It is the number of users who enrolled in the free trial, divided by the number of users who clicked the "start free trial" button. After having clicked the "start free trial" button, a popup page appears for the experiment group asking for the amount of time that the student can devote to the course and then makes a suggestion whether the student should enroll for the course or continue with the free courseware. In the experiment group the user can make a decision, if he has enough time to devote for the course, he enrolls for the course or else he continues with the free courseware without enrolling. On the other hand, for the control group no popup page appears, thus the user regardless of his time availability, he enrolls for the course. Hence, the gross conversion might be more in the control group and can be used as an evaluation metric to check if the experiment makes a significant difference in the enrollment. Therefore we should tell by this metric if we significantly reduced the number who left the free trial early. Decreased enrollment is what we are looking for.

### Retention:

It is the number of user-ids to remain enrolled for 14 days trial period and make their first payment, divided by the number of users who enrolled in the free trial. As explained for the above metric, since the users in the experiment group are made aware about the minimum amount of time they must devote for the course, they make a decision to enroll for the course or continue with the free courseware. The retention ratio might be high for the experiment group, since relatively fewer users would enroll based on their availability of time and most of them would remain enrolled for a 14 day period and make their payment. But for the control group many users enroll for the courseware, but since not many of them have the required amount of time available, they cancel their enrollment in the free trial period and hence the retention ratio will be low. Thus, it can be used as an Evaluation metric and for the same reason it cannot be used as an Invariant metric. We should tell by this metric if we significantly reduced the number who left the free trial early.

### Net Conversion:

It is the number of user-ids to remain enrolled for the 14 days trial period and make their first payment, divided by the number of users who clicked the “start free trial” button. For the experiment group, the users are made aware about their minimum time availability by the free trial screener page and hence they can make a decision to remain enrolled for the course after the 14 day trail. But for the control group, the users are not aware of their minimum time availability and might not cancel their subscription before the 14 day trial and make a payment. Thus, it can be used as an Evaluation metric and for the same reason it cannot be used as an Invariant metric. By this Evaluation metric we should tell if we did not significantly reduce the number who continue past the free trial.

### Non-usable metrics

#### Number of user-ids:

It is the number of users who enroll in the free trial. Since the unit of diversion is the cookie, and students are tracked by user-id after they enroll in the free trial, number of user-ids is not a good metric to be used either as an Evaluation or as an Invariant metric. It wouldn't notice any change between the control and experiment group.

#### Click-through-probability:

It is the number of unique cookies to click the "start free trial" button divided by number of unique cookies to view the course overview page. Since number of clicks and number of cookies are already used as Invariant metrics and click-through-probability is a fraction of those two, it could be redundant to use it as an Invariant metric as well. In addition, both number of clicks and number of cookies are equally distributed across the control and experiment group. Thus the fraction of those cannot be used as an Evaluation metric.

To launch the experiment we must observe a decrease in gross conversion, an increase in retentions and a stability in net conversion.

## Measuring Standard Deviation

Evaluation Metric	Standard Deviation
Gross Conversion	0.0202
Retention	0.0549
Net Conversion	0.0156

For Gross Conversion and Net Conversion, the Unit of Diversion is equal to the Unit of Analysis. After plotting the distribution of the difference between the metrics for control and experiment group, it seemed approximately normal. Thus the analytical estimate would be similar to the empirical variability.

For Retention the Unit of Diversion is not equal to the Unit of Analysis which is the “number of users who enrolled in the courseware”. Hence the empirical variability may be different from the analytical estimate, thus we perform both an analytical and empirical estimate for this metric.

## Sizing

### Number of Samples vs. Power

We will use Bonferroni correction during the analysis phase. The [alpha\\_overall](#) will be 0.05.

Thus, the [alpha\\_individual](#) will be  $0.05 / n$ , where  $n$  is the number of Evaluation metrics.

After calculating the number of pageviews for each metric, we get the results on the table below.

Evaluation Metric	Number of Pageviews
Gross Conversion	825,350
Retention	7,054,424
Net Conversion	875,000

In order to perform the experiment appropriately we will need the greatest number of pageviews of the three Evaluation metrics. Consequently, we will need 7,054,424 pageviews.

### Duration vs. Exposure

Knowing that we don't need to perform any other experiment for the time being, I would divert the entire amount of traffic of Udacity for this experiment. In addition, knowing that this experiment is made for the good of the users by providing them with information about the minimum amount of time they need to devote to the courses, diverting the entire traffic will not

harm Udacity. Even if the experiment doesn't work, it could easily revert back to the previous state for the experiment group without any trouble. Hence, it is not too risky for the company.

The total number of days needed to perform the experiment is calculated by the following fraction:

$7,054,424 \text{ pageviews} / 40,000 \text{ pageviews per day} = \text{approximately } 176 \text{ days}$

After inspecting that the number of days is considerably high, it may take a lot of time to perform the experiment which may not be feasible for the company, so we need to rethink the metrics to perform the experiment. Thus, we should **exclude the Retention** and **choose only Gross conversion and Net conversion** as the Evaluation Metrics.

The total number of days needed to perform the experiment after excluding Retention from the Evaluation metrics is calculated by the following fraction:

$875,000 \text{ pageviews} / 40,000 \text{ pageviews per day} = \text{approximately } 22 \text{ days}$

With the changed metrics we can perform the experiment within 22 days which is practically and financially feasible for the company.

## Experiment Analysis

### Sanity Checks

Number of Cookies:

Control Group	Experiment Group	Total
345,543	344,660	690,203

Probability of a cookie fallen into the control or the experiment group = 0.5

Standard Error (SE) =  $\sqrt{(0.5 \cdot 0.5 / (345,543 + 344,660))}$  = 0.0006

Margin of error (m) = SE \* 1.96 = 0.0011

Confidence Interval =  $[0.5 - m, 0.5 + m]$  = [0.4989, 0.5011]

Observed Value =  $344,660 / 690,203$  = 0.5006

It is within the Confidence Interval and **passes the Sanity Check**

---

Number of Clicks:

Control Group	Experiment Group	Total
---------------	------------------	-------

28,378	28,325	56,703
--------	--------	--------

Probability of a click fallen into the control or the experiment group = 0.5

Standard Error (SE) =  $\sqrt{(0.5 \cdot 0.5 / (28,378 + 28,325))}$  = 0.0021

Margin of error (m) = SE \* 1.96 = 0.0041

Confidence Interval = [0.5 - m, 0.5 + m] = [0.4959, 0.5041]

Observed Value = 28,325 / 56,703 = 0.50046

It is within the Confidence Interval and [passes the Sanity Check](#)

## Result Analysis

### Effect Size Tests

Earlier we mentioned that using Bonferroni correction for multiple metrics, the alpha\_overall will be 0.05. Thus, the alpha\_individual will be 0.05 / n , where n is the number of Evaluation metrics. In our case, alpha\_overall = 0.05 and alpha\_individual = 0.05 / 2 = 0.025 with z-score = 2.24.

### Gross Conversion:

	Control Group	Experiment Group
Clicks	17,293	17,260
Enrollments	3,785	3,423
Gross Conversion	0.2188746892	0.1983198146

d_min : minimum practical significance	±0.01
SE	0.004371675385
m = SE * 2.24	0.009792552863
Difference	- 0.02055487458
Confidence Interval	[ - 0.03034742744, - 0.01076232172]
Statistically Significant	<a href="#">Yes</a> , since CI does not contain zero
Practically Significant	<a href="#">Yes</a> , since CI does not contain the d_min value

### Net Conversion:

	Control Group	Experiment Group
Clicks	17,293	17,260
Enrollments	2,033	1,945
Gross Conversion	0.1175620193	0.1126882966

d_min : minimum practical significance	±0.0075
SE	0.003434133513
m = SE * 2.24	0.007692459069
Difference	- 0.00487722675
Confidence Interval	[- 0.01256618174, 0.002818736394]
Statistically Significant	No, since CI contains zero within it
Practically Significant	No, since CI contains the d_min value within it

## Sign Tests

### Gross Conversion:

alpha_individual	0.0025
Number of successes	4
Number of trials	23
Probability	0.5
Two-tailed p-value	0.0026

Since, it is less than the given alpha level it is statistically significant. It agrees with the hypothesis test for the Gross Conversion which was statistically and practically significant.

---

### Net Conversion:

alpha_individual	0.0025
Number of successes	10
Number of trials	23
Probability	0.5

Two-tailed p-value	0.6776
--------------------	--------

By inspecting the above table one can see that the result is no less than the  $\alpha_{\text{individual}}$  and therefore it is not statistically significant. Thus it comes in agreement with the hypothesis test for the Net Conversion which was also not statistically significant.

## Summary

We have used Bonferroni correction for the testing, for the reason that we are using multiple metrics for experiment. The problem of multiplicity arises from the fact that as we increase the number of hypotheses being tested, we also increase the likelihood of a rare event, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e. make a Type I error). The Bonferroni correction is based on the idea that if an experimenter is testing “m” hypotheses, then one way of maintaining the familywise error rate (FWER) is to test each individual hypothesis at a statistical significance level of “1/m” times what it would be if only one hypothesis was tested. In our case,  $m = 2$ ,  $\alpha_{\text{overall}} = 0.05$  and by using Bonferroni correction we get  $\alpha_{\text{individual}} = 0.025$ .

Consequently, using a Bonferroni correction does not seem correct. In our cases a correction is required but the Bonferroni correction would be too conservative because there is no great justification for controlling the type I errors and by reducing power increases the chance of a false negative. If a correction is required but the original Bonferroni procedure is regarded as too conservative, then a possible alternative is to use the Bonferroni-Holm or Hochberg methods.

## Recommendation

In the experiment above we performed two hypothesis tests, one was Gross Conversion and the other was Net Conversion.

Gross Conversion:

It is the ratio of the number of users enrolling in the course to the number of users who clicked the “start now” button. The hypothesis test was both statistically and practically significant, showing that the popup page recommending the user about the minimum dedication time has a positive effect on the experiment group reducing the number of users enrolling for the free trial option. This is good for Udacity as the coaches might concentrate more on less number of students and their probability of enrollment after the 14 days trial would increase. The sign test was also statistically significant giving us more confidence to recommend the launch of the experiment.

Net conversion:

It is the ratio of the number of users who remain enrolled for the 14 day trail and make their first payment to the number of users who clicked the “start free trial” button. The hypothesis test and sign test for this experiment was both statistically and practically not significant. The reason for insignificance might be:

The popup page making the recommendation about the minimum dedication time appears after clicking the “start free trial” button for the experiment group but not for the control group. The users in the experiment group receive the suggestion and decide to

either enroll for the free trial or opt out for the open courseware. The users in the control group on the other hand might not have received the suggestion and would have enrolled for the free trial period, but within the 14 days trial they become aware of their availability of the time and might not choose to make the payment and opt out for the courseware within the 14 day free trial without making the payment.

Thus, the number of users actually making the payment in both groups might not be statistically significant, which we observed in our hypothesis test. Even if the experiment doesn't make a difference in the Net Conversion, the confidence interval does include the negative of the practical significance boundary and it still makes a significant difference in the Gross Conversion which is good for Udacity. It's possible that this number went down by an amount that would matter to the business. This risk isn't worth taking.

## Follow-Up Experiment

In order to reduce the number of dissatisfied students we can perform the following experiment:

We could design a course which is based on the prerequisites for the Nanodegree that is doable based on the required level of the student in the 14 day trial period with around 5hrs/week involvement. This attempt will help Udacity in two ways:

1. It will make the students aware of the prerequisite level of knowledge that they need to have in order to successfully complete the Nanodegree
2. It will help the students determine the actual amount of time they should devote to the courseware and if it is sufficient to complete the Nanodegree within time. The experiment group on clicking the "start free trial" will be asked to take the pre-course before moving further in the core courses and suggest successfully complete the pre-course during the 14 days trial. This will make them the ideal candidates for the Nanodegrees.

Hypothesis:

The number of students successfully completing the pre-course in the 14 days trial period are more likely to remain enrolled in the Nanodegree and make their first payment.

Initial Unit of Diversion:

User-ids

Invariant Metrics:

1. Number of user-ids: Since the Unit of diversion is user-id based, thus the users are evenly split between the control and experiment groups based on the user-id, thus it is a good Invariant Metric.

Evaluation Metrics:

1. Retention: It is the number of user-ids(students) to remain enrolled for 14 days trial period and make their first payment, divided by the number of users who enrolled in the free trial. Since the users in the experiment group are aware about the minimum amount



of time they must devote for the course, they make a decision to enroll for the course or continue with the free courseware. The retention ratio might be high for the experiment group, since relatively fewer users would enroll based on their availability of time and most of them would remain enrolled for a 14 day period and make their payment. But for the control group many users enroll for the courseware, but since not many of them have the required amount of time available, they cancel their enrollment in the free trial period and hence the retention ratio will be low. Thus, it can be used as an Evaluation metric.