

# Design and Analyze A/B Testing

By Carlos Hernandez

## Experiment Design Metric Choice

### Invariant Metrics

- Number of cookies, Number of clicks.

### Evaluation Metrics

- Gross Conversion, Retention, Net conversion.

### Explanation

**Number of cookies.** Good invariant metric because visits are independent from the experiment, unit of diversion: cookies.

**Number of user-ids.** Not a good invariant metric because the users that enroll the free trial are dependent of the experiment, see different values in control and experiment groups is expected. It could be used poorly as evaluation metric because is redundant with other metrics like gross conversion, which gives a better result reducing the number of students to continue past the free trial.

**Number of clicks.** Good invariant metric because clicks occur before users see the experiment, and are not dependent.

**Click-through-probability.** Good invariant metric because clicks occur before users see the experiment, therefore there is not dependency, but I think number of cookies and number of clicks are good enough.

**Gross Conversion.** Good evaluation metric because it depends on the experiment directly, it can be used to check if the experiment makes a significant difference in the enrollment process.

**Retention.** Good evaluation metric, because it depends directly of the experiment and shows a financial outcome of the tested change.

**Net conversion.** This is a good evaluation metric, is directly dependent on the effect of the experiment and will help determine if the ratio of user making a payment and total users that see the experiment will increase.

In order to launch the experiment we expect the Gross Conversion to have a decrease in practical significance, Net Conversion to have an increase in statistical significance. We will look at this two metrics because the first one tell us if we are going to reduce cost if the change is introduced and the second one will tell us how the change affects revenue.

## Measuring Standard Deviation

### Gross Conversion

SE = 0.0202

### Retention

SE = 0.05495

### Net Conversion

SE = 0.0156

Gross Conversion and Net Conversion has number of cookies as denominator, which is the unit of diversion that is equal to unit of analysis, so the analytical estimate can be similar to the empirical.

Retention has number of user-ids enrolled as denominator, which is not the unit of diversion and different from the unit of analysis, so the analytical estimate is not equal to the empirical.

For these reasons we can use Gross Conversion and Net Conversion for calculating the analytical estimate.

## Sizing

### Number of Samples vs. Power

I am not going to use Bonferroni correction on analysis phase because this correction will be too conservative.

According to calculation, 685,275 pageviews will be needed in the experiment to give it power with Gross Conversion and Net Conversion metrics, this value includes control and experiment groups.

### Duration vs. Exposure

With traffic of 40,000 unique cookies to view page per day, I suggest to send 60% of the traffic (34,000) to the experiment, so it could take ~29 days to complete the experiment.

I think this is a reasonable fraction that will not affect the existing users/customers and will impact new enrollments. As we saw on A/B Testing course, directing 100% of the traffic could cause bugs and/or issues and is not recommended.

## Experiment Analysis

### Sanity Checks

#### Number of cookies.

Confidence Interval: [0.4988, 0.5012]

Observed Value: 0.5006

PASS

#### Number of clicks.

Confidence Interval: [0.4959, 0.5041]

Observed Value: 0.5005

PASS

## Result Analysis

### Effect Size Tests

#### Gross Conversion

	Control Group	Experiment Group
<b>Clicks</b>	17,293	17,260
<b>Enrollment</b>	3,785	3,423
<b>Gross Conversion</b>	0.2189	0.1983

Confidence Interval: [-0.0291, -0.0120]

Statistically significant

Practically significant

#### Net Conversion

	Control Group	Experiment Group
<b>Clicks</b>	17,293	17,260
<b>Payments</b>	2,033	1,945
<b>Net Conversion</b>	0.1176	0.1127

Confidence Interval: [-0.0116, 0.0019]

Not statistically significant

Not practically significant

## Sign Tests

### Gross Conversion

Alpha:	0.025	
Number of success:	4	
Number of trials:	23	
Probability:	0.5	
p-value:	0.0026	statistically significant

### Net Conversion

Alpha:	0.025	
Number of success:	10	
Number of trials:	23	
Probability:	0.5	
p-value:	0.6776	not statistically significant

## Summary

Bonferroni correction was not necessary because the metrics in the test has high variance and in this case, to launch the experiment we are using “all” the metrics, if correction is applied, it would be too conservative. In other case, if we use “any” metrics, Bonferroni correction surely is needed.

## Recommendation

Gross Conversion resulted negative and practically significant. This is a good result because costs can be reduced by discarding trial sign ups that will not convert.

Net Conversion resulted not statistically significant and not practically significant with negative numbers on interval confidence. This is a bad result because this metric could get down and have a negative impact on Udacity business.

I would recommend improve the experiment design and analyze the results before launch the new feature, if better results are obtained, it could be launched.

## Follow-Up Experiment

Udacity makes a great job including top new tech courses at a convenient price democratizing excellent online education, the quality of most of them exceed the expectations that students have, however, after taking nanodegrees and reviewing feedback on forums, there are some courses with lower quality than others, maybe including a 5 star ranking for each course and nanodegree Udacity could get more users engaged with the programs. New courses and

programs could be tested applying A/B testing to determine if a course has good quality or if it will be in demand by students before launching.

**Null hypothesis:** by creating a rating on each course, it will not increase Retention by a practically significant amount.

**Unit of diversion:** User-ids, users enrolled on courses with >4 stars will be tracked by user-id.

**Invariant metric:** Number of user-ids, because the users will sign up for free trial after reviewing high rated courses.

**Evaluation metric:** Retention, because it will show an increase in revenue in case that high rated courses have been seen and taken by students, and if this metric is positive and practically significant.