P7 - Design an A/B Test

Student: Peter Carsten Petersen

Experiment Design

Metric Choice

Metric	Invariant	Evaluation
Number of Cookies	**YES** Data is from number of Unique cookies viewing the course overview page, i.e. data captured before click on “Start Free Trial” button which prompts the change in our experiment. So all else equal this Metric should not change due to experiment, and therefore an ideal Invariant metric.	**NO** Data captured before click on “Start Free Trial” button and the subsequent event which is the topic of the experiment, and therefore this metric not expected to change due to experiment. Consequently Metric cannot be used for Evaluation.
Number of User Id’s	**NO** Data is collected after click on “Free Trial button” and subsequent enrollment. Therefore highly influenced by the experiment and likely to change and can therefore not be used as Invariant.	**NO** Number of User Id’s could potentially be used for Evaluation Metric, to measure the first part of our hypothesis “Reducing the number of frustrated students…..”, but since “Number of User Id’s” is a simple count we cannot expect it to be Normalized, and we will choose “Gross Conversion” instead, as the better option for Evaluation.
Number of Clicks	**YES** Data is from number of Unique cookies clicking the “Start Free Trial” button which prompts the change for the experiment. So all else equal this should not change due to experiment, and therefore ideal Invariant.	**NO** Data captured on click on “Start Free Trial” button which in turn prompts the change in the experiment. But the click event in itself will not be influenced by the change and therefore cannot be used for Evaluation.
Click Through Probability	**YES** This data is the Number of Clicks divided by Cookies and takes place before Change Experiment. So all else equal this should not change due to	**NO** Data captured on click on “Start Free Trial” button which in turn prompts the change in the experiment. But the click through probability on “Start Free Trial”

	experiment, and therefore ideal Invariant.	button in itself will not be influenced by the change and therefore cannot be used for Evaluation.
Gross Conversion **(dmin = 0.01)**	**NO** Data is captured after click on “Start Free Trial” and consequently for our Experiment group the Test of adding a prompt for considering available time for course, before possible Enrolling. We will therefore expect a change in this Metric, which therefore cannot be used as Invariant Metric.	**YES** Data is captured after click on “Start Free Trial” and consequently for our Experiment group the Test of adding a prompt for considering available time for course, before possible Enrolling. This is therefore the ideal Metric for Evaluating part of the Hypothesis of the Test. “Reducing the number of frustrated students…..”
Retention	**NO** Data captured after click on “Start Free Trial” and for Control group in our test, after prompt for considering available time. We will therefore expect a change in this Metric, which therefore cannot be used as Invariant Metric.	**NO** Could potentially have been used for Evaluation of the 2nd part of our Hypothesis “without significantly reducing the number of students to continue past the free trial……”. but since the denominator is “User Id’s to complete checkout, we will be missing the Students who did not enroll in the analysis and therefore Net Conversion is the better Evaluation Metric.
Net Conversion **(dmin = 0.0075)**	**NO** Data is captured after 14 day Trial Period and consequently for our Experiment group in the Test after adding a prompt for considering available time for course, before possible Enrolling. We will therefore expect a change in this Metric, and therefore cannot be used as Invariant Metric.	**YES** Data is captured after 14 day Free Trial period, and consequently for our Experiment group the Test also after adding a prompt for considering available time for course, before possible Enrolling. This is therefore the ideal Metric for Evaluating second part of the Hypothesis of the Test. “without significantly reducing the number of students to continue past the free trial……”

Experiment Launch: I would expect the result of this experiment, to show a negative effect on the **Gross Conversion,** as some potential students will be discouraged by the prompt for considering necessary time needed to complete the course.

For **Net Conversion** I would also expect some negative effect, but to a lesser extent than **Gross Conversion** as the pool of discouraged students from **Gross Conversion** were already likely candidates to leave before end of Free Trial.

In order to recommend a launch of the change post experiment based on the hypothesis, I will be looking for a change in Gross Conversion equal or greater than the dmin of 0.01 and a Net Conversion with a positive change or negative less than dmin of 0.0075.

Measuring Standard Deviation

Evaluation Metric	Standard Deviation
Gross Conversion	0.0202
Net Conversion	0.0156

Analytical vs. Empirical:

For both metrics Unit of analysis and Unit of Diversion are Unique cookies, and since this is provides us with a match between numerator and denominator, we can go ahead and use the analytical estimate for Analysis. Consequently an Empirical estimate will not be needed.

Sizing

Number of Samples vs. Power

Description	Use method Yes/No	Reason
Bonferroni Method	No	Bonferroni Method not used as both Evaluation Metrics need to meet expectations. Using Bonferroni Method would increase risk of a False Negative, whereas we are trying to minimise this exact risk.

Evaluation Metric	Parameters	Pageviews Needed for analysis
Gross Conversion	prob. 20.63% / α = 0.05 / β = 0.2 / dmin = 1%	647,325
Net Conversion	prob. 10.93% / α = 0.05 / β = 0.2 / dmin = 0.75%	685,275 ✓

For pageviews needed in the A/B test, we must choose the highest number calculated on each of the evaluation metrics. I.e. 685,275 pageviews will be needed.

Duration vs. Exposure

Pageviews Needed	Fraction of traffic diverted	Length of experiment in Days
685,275	1.0	18

Reason and Risk: The Fraction of traffic diverted and consequent number of days needed, is based on a review of the 4 Main Principles of A/B Testing:

1. **Risk:** The risk in this particular test is viewed as “Below Minimal Risk”, as the simple adding of a “question box” in the enrollment process cannot do any harm.
2. **Benefits:** The aim of the Test is to decrease the Frustration of some students and allow for more coaching of remaining students. A clear aim at increasing quality and hence a pass of the Benefits principle.
3. **Alternatives:** Students will have some alternatives to Udacity, but when clicking the “Start Free Trial”, a simple “question box” will not likely prompt a search for an alternative, and therefore this Principle is viewed as low risk.
4. **Data Sensitivity:** This experiment, while collecting sensitive personal and financial information when enrolling, is not collecting any new or added sensitive data which would not be collected otherwise.

Based on review of the 4 Main Principles the Experiment is overall viewed as low risk, and I recommend Test be run on entire traffic, in order to minimize length of experiment.

Experiment Analysis

Sanity Checks

Invariant Metric	Lower Bound	Upper Bound	Observed	Passed
Number of Cookies	0.4988	0.5012	0.5006	✓
Number of Clicks	0.4959	0.5041	0.5005	✓
Click Through Probability	0.0812	0.0830	0.0822	✓

All Sanity Checks on Invariant Metrics are passed, and point towards a successfully run test.

Result Analysis

Effect Size Tests

Evaluation Metric	Lower Bound	Upper Bound	dmin	Statistical Significant	Practical Significant
Gross Conversion	-0.0291	-0.0120	+/- 0.01	✓	✓
Net Conversion	-0.0116	0.0019	+/-0.0075	%	%

Sign Tests

Evaluation Metric	Days where experiment > control	Total Days	two-tailed p value	Statistical Significant
Gross Conversion	4	23	0.0026	✓
Net Conversion	10	23	0.6776	%

Summary

Bonferroni Method not used as both Evaluation Metrics “Gross Conversion” and “Net Conversion” need to match expectations. Therefore we are concerned with the Risk of a Type II error i.e. failing to detect an effect that is present. Bonferroni Method will increase the confidence interval, at the risk of resulting in a False Negative.

Bonferroni Method is most effective when dealing with Metrics where only 1 of many needs to match expectations and where the aim is to reduce the risk of a Type I error i.e. detecting an effect that is not present. In this case the Bonferroni Method will increase the confidence interval and results will have to be very strong too meet expectations, reducing risk of false positives.

No discrepancies between result of Effect Size and Sign tests.

Recommendation

Recap of goal of the tested change: *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.*

Based on the A/B Test result I would make a recommendation **not to implement this change.**

The **Gross Conversion** is showing a both statistical and Practical Significant change in the negative direction with a confidence interval between -0.012 and -0.029, this therefore meeting expectations on the “reducing the number of frustrated students….” part of hypothesis.

The **Net Conversion** however is not showing the same clear test result. A difference of -0.0049 between experiment and control is within the acceptable change, but the 95% confidence interval is between -0.0116 and 0.0019, and therefore a higher negative risk that if launched could result in a level of **Net Conversion** above the -0.0075 dmin level.

Since we require both Metrics to be within expectations we cannot launch based on this test result.

Follow-Up Experiment

In order to increase the **Net Conversion** of students enrolling in Free Trial, I would recommend testing, **post Free Trial Enrollment,** an added course segment consisting of a series of small explanatory Video’s outlining for each segment of course:

1. High Level description of specific course segment and presentation of instructors.
2. Description and visualization of tools mastered post each segment completion.
3. Real life examples of course segments.
4. A maximum duration of 3-5 minutes per video.

The Hypothesis would be an increase of **Retention** as students become more aware of the interesting path ahead, and the benefits of completing and acquiring new competencies.

Invariant Metric: Will be the **Number of User ID’s** (Number of users to enroll in the Free Trial), which will also be used as our unit of Diversion.

Evaluation Metrics: Will in this case only be the **Retention** (user id’s staying on after 14 day Free trial Boundary divided by number of User ID’s to complete checkout) **,** as we are only testing the rate of students who stay on after end of Free Trial.

dmin should be calculated before Testing, as a product of minimum uplift needed in **Retention** in order to recover cost of producing video content within a reasonable timeframe. This calculated **dmin** should furthermore pass a simple feasibility consideration before possible testing is initiated.

Resources: www.wikipedia.org / stats.stackexchange.com
www.evanmiller.org/ab-testing/sample-size.html / graphpad.com/quickcalcs/binomial1/