Final Project of A/B Testing by Dogan Askan

Experiment Design

Metric Choice

Invariant Metrics

- Number of cookies
- Number of clicks

Evaluation Metrics

- Gross conversion
- Retention
- Net conversion

Number of cookies: Users are separated for the experiment after seeing the overview page. This metric is expected to be similar for both experiment and control groups so I chose this as an invariant metric.
Number of user-ids: This can be used as an evaluation metric because it would track the first part of the hypothesis, namely that we will reduce the number of students to continue past the free trial. However, since there are better metrics for evaluation I didn't choose this.
Number of clicks: Like the **Number of cookies,** users are separated for the experiment just after clicking the "Start free trial" button. This metric is also expected to be similar for both experiment and control groups so I chose this as an invariant metric, too.
Click-through-probability: Since this is the function of both **Number of cookies** and **Number of clicks** and both are chosen as invariant metrics and we have already enough number of invariant metrics, it is unnecessary to chose this as a metric even though it may be a better invariant metric compared to **Number of clicks**.
Gross conversion: Since it is directly related with the conversion and takes place after the intervention, it is expected to be lower after the change, so this is also suitable for evaluation metrics.
Retention: There are two parts in the hypothesis. In the second part, **Retention** should not change significantly, a small amount of reducing in this may be acceptable. It also takes place after the intervention. Thus we need to evaluate this metric.
Net conversion: This is the essential metric we need to track in the hypothesis. More people are expected to stay in the enrollment after the change. Since the hypothesis indicates that number of giving up students before 14 days free trial ends decreases after the change, I also chose this as an evaluation metric.

In order to launch the experiment, and if we continue to keep all evaluation metrics, **Retention** and **Net conversion** should not decrease significantly and the difference needs to be greater than dmin= 0.01 in **Gross conversion** negatively.

Measuring Standard Deviation

Standard Deviations of Evaluation Metrics
Gross conversion = 0.0202
Retention = 0.0549
Net conversion = 0.0156

Number of cookies is in denominator for **Net Conversion** and **Gross conversion**, and it is unit of diversion. Since the unit of diversion is equal to unit of analysis, it indicates that the analytical estimate would be comparable to the empirical variability.

For **Retention**, the denominator is not same as unit of diversion which means analytical estimate would not be comparable to the empirical variability. Having said that, if I had time I would want to collect an empirical estimate of the variability for this metric.

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.

Sizing

Number of Samples vs. Power
I will not use the Bonferroni correction during my analysis.

	sample size per group	exp + cont	per view	pageview needed
Gross conversion	25835	51670	0.08	645875
Retention	39115	78230	0.0165	4741212
Net conversion	27411	54822	0.08	685275

Duration vs. Exposure
The risk is not more than minimal risk that the students won't encounter a greater risk due to the change than the risk they encounter during their normal life. So I would divert 100% of traffic to this experiment. In this scenario, we 119 days to run the experiment by using **Retention** as one of the evaluation metrics and by considering its pageviews needed. This is an incredibly huge number, nobody wants to wait that long and during this period many things may change. Thus, by eliminating **Retention** and keeping other factors same, we can continue with second highest number of pageview needed, that metric is **Net conversion.** So, we can reduce the duration to 18 days by considering **Net conversion**'s pageviews needed. The experiment is not risky, we should be fine.

Experiment Analysis

Sanity Checks

	Pageviews	Clicks
Expected	0.5	0.5
Observed	0.500640	0.500467
SE	0.000602	0.002100
z(a=0.05)	1.96	1.96
m	0.001180	0.004116
upper limit	0.501180	0.504116
lower limit	0.498820	0.495884
pass	YES	YES

Result Analysis

Effect Size Tests

	Net conversion		Gross Conversion	
	Control	Experiment	Control	Experiment
X	2033	1945	3785	3423
N	17293	17260	17293	17260
p_pooled	0.1151274853		0.2086070674	
SE_pooled	0.003434133513		0.004371675385	
d	-0.004873722675		-0.02055487458	
z(a=0.05)	1.96		1.96	
m	0.006730901685		0.008568483755	
Upper limit	0.001857179011		-0.01198639083	
Lower limit	-0.01160462436		-0.02912335834	
Statistically	Not significant		Significant	
Practically	Not significant		Significant	

Sign Tests

	Net conversion	Gross conversion
Succes	10	4
Trial	23	23
Probability	0.5	0.5
P-value	0.6776	0.0026
alpha	0.05	0.05
Significant	No	Yes

Summary
I did not use Bonferroni correction because all tests are required to be significant, not some are enough. We are not able to launch the feature if the change in at least one metric doesn't pass the criteria. In our case, we need both metrics to pass the criteria and the change in them are significant.

Effect size and sign test show that the change affects **Gross Conversion** but not **Net conversion** significantly**.**

Recommendation

We tested the hypothesis that adding the commitment screen before enroll reduce the student number who left before free trial ends without significantly reducing the number of students to continue past the free trial and eventually complete the course. The result in **Gross Conversion** is significant and negative which means the change indeed reduce the number of frustrated students who left the free trail because they didn't have enough time. However, since the result in **Net conversion** is not significant and the confidence interval of the **Net conversion** does include the negative of the practical significance boundary. Thus, it's possible that this number went down by an amount that would matter to the business and this is not an acceptable risk in order to launch. This may indicate that the change may also demotivate the students who have actually enough time to devote in addition to the students who left the free trail because they didn't have enough time. Consequently, launching this feature may have some negative effects and it would be a risky decision. Considering additional experiments or other design choices is a good idea.

Follow-Up Experiment

I think the message in the current experiment may be slightly demotivator for some students and it looks the main problem is so many students left in the free trial, almost half of them.For this reason, I propose a motivator message set in the follow-up experiment. Messages indicating some successful stories of graduated students would be helpful. For example, “John Doe is now working for this great company after committed 6 hours/week on average at Udacity.”. 14 different messages from 14 different succesfully graduated students who got his/her job would be shown whenever a student start studying, not more than 1 daily. This kind of message can motivate student and transmit the commitment messages as well.

So the hypothesis might both motivate and set clearer expectations for students, thus more students would continue after free trial without changing the enrolled students number. If this hypothesis hold true, Udacity turns more students who left in free trial into the students continue after free trial.

In this case, unit of diversion would be **user-ids** since the students already got their ids because they already enrolled.

Invariant metric would be **Number of user-ids** due to the similar reasoning in the current experiment.

Evaluation metric would be the ratio of number of user-ids to remain enrolled past the 14-day boundary over number of users who enroll in the free trial called **ratio of continued students.** By using this, we can get clearer result how successful we are after the experiment.

Resources

http://graphpad.com/quickcalcs/binomial2/
http://www.evanmiller.org/ab-testing/sample-size.html
http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full