Experiment Design

Metric Choice

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: This metric is invariant metric. Because we want to constant the number of sources. The number of cookies as the start of the funnel. It should be controlled in the experiment.

Number of user-ids: Neither invariant metric nor evaluation metric. This metric can be affected by this experiment. It doesn't make sense as invariant metric due to the experiment changes. It could be the potential evaluation metric. But not a good evaluation metric. Because even the cookies are assigned randomly. The number for each group may not same. So the use-ids can be affected by the previous step. Even the number from previous is equal. It is hard to evaluate statistical significance.

Number of clicks: Neither invariant metric nor evaluation metric. The reason why it should not be invariant metric is this number of click can be affected by the trial screener. If this hypothesis is true. The number can be impacted significantly from the previous step of the funnel. The reason why it should not be the evaluation metric is same to the # user-ids, it is hard to measure the different.

Click through rate: This metric is invariant metric. As mentioned in the overview, "without significantly reducing the number of students to continue past the free trial". So it should be invariant.

Gross conversion: This metric is chosen as an evaluation metric as we are hoping to decrease the people who likely to drop from doing the checkout. It should reduce the conversion rate for the experiment group.

Retention: Invariant metric. After decreasing the people who are likely to drop from doing the checkout. As instruction, it should be nothing happen to this metric. This metric is just one nice to have. Even though it can also measure the user remain enrolled for the past the 14-days free trial significantly. But the denominator is too small. It will take too long time to get the result if we use this metric as the evaluation metric.

Net conversion: Evaluation metric. It is also able to measure the rate of people enrolled pass the 14-days free trial. And this metric is better than retention due to the duration. This also should be nothing happen or not decrease too much to the experiment group.

Measuring Standard Deviation

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to 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.

Because the unit of analysis for the metrics below are cookies. The empirical variability and analytical variability should be consistent. The analytical variability is easy to calculate. If the unit of analysis and the unit of diversion in the evaluation metrics are different. We should use the empirical variability.

Gross conversion:
P = 0.206
N = 400
SD = sqrt(P * (1-P)/N) = **0.0202**

Net conversion:
P = 0.1093
N = 400
SD = sqrt(P*(1-P)/N) = **0.0156**

Sizing

Number of Samples vs. Power

Pageview Number: **685325**

We need to use the online calculator to get the number of pageviews for each metric individually and choose the largest value as the amount need.

For this time not use Bonferroni correction, because it should be coefficient for these evaluation metrics.

Gross conversion sample size:
Probability of enrolling/given click:0.20625
d_min = 1%
sample size: 25830
Total page view: 25830/(3200/40000) * 2 = 645750

Net conversion sample size:
Probability of payment/ given click:0.1093125
d_min = .75%
sample size: 27413
Total page view: 27413(3200/40000) * 2 = 685325

Duration vs. Exposure

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

If we use the 685325 as pageviews. According to the traffic per day from baseline spreadsheet. We need to take 18 days to get the result even though using 1 as the fraction of the traffic. But if we use 1 as the fraction. It may impact too many people and any bug or negative effect will impact all user. The risk of the experiment of this test many distress the student who don't have

enough time for the class and make them feel they are not valuable and frustrate. It has the possibility of reducing the income of the student who gives up enroll.
We can also use 0.5 as the traffic, it will take 36 days to get the result. Lower traffic can make several AB test at the same time.

Experiment Analysis

Sanity Checks

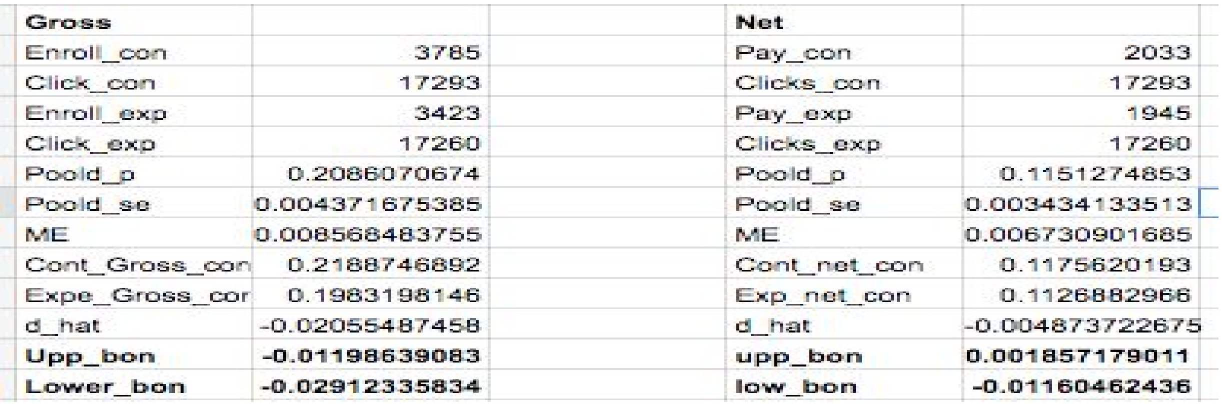
For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

Number of cookies:
Ncont = 345543
Nexp = 344660
SD = sqrt(0.5*0.5/(Ncont+Nexp))
ME = 1.96*SD
Upper_bound = 0.5 + ME = 0.50118
Lower_bound = 0.5 - ME = 0.49882
Observed = Ncon/(Ncon + Nexp) = 0.50064
Pass the sanity check
CTR:
p_pool = (Xcon+Xexp)/(Ncon+Nexp) = 0.08215
se_pool = sqrt(p*(1-p)*(1/Ncon+1/Nexp)) = 0.000661
me = 1.96*se_pool = 0.001296
d^ = X_con/Ncon-X_exp/NEexp = 0.00005627
Pass the sanity check

If the sanity check didn't pass.
Maybe something wrong with the classifier which assigns one of the groups too many.

Result Analysis

Effect Size Tests



Gross			**Net**	
Enroll_con	3785		Pay_con	2033
Click_con	17293		Clicks_con	17293
Enroll_exp	3423		Pay_exp	1945
Click_exp	17260		Clicks_exp	17260
Poold_p	0.2086070674		Poold_p	0.1151274853
Poold_se	0.004371675385		Poold_se	0.003434133513
ME	0.008568483755		ME	0.006730901685
Cont_Gross_con	0.2188746892		Cont_net_con	0.1175620193
Expe_Gross_cor	0.1983198146		Exp_net_con	0.1126882966
d_hat	-0.02055487458		d_hat	-0.004873722675
Upp_bon	**-0.01198639083**		**upp_bon**	**0.001857179011**
Lower_bon	**-0.02912335834**		**low_bon**	**-0.01160462436**

	Statistical significance	Practical significance
Gross	Y	Y
Net	N	N

Sign Tests

Number of sign test for each of evaluation metrics according to the online calculator:
Gross : 4/23
P-value = 0.0026
Retention : 13/23
P-value = 0.6776
Net : 10/23
P-value: 0.6776

The gross is statistical significant but the Net convention is not.

If we do not pass the sign test, we show trying to find the most noticeable different. We might break down to the different platform, different days of a week. This not only helps to find the bugs, it also might give a new hypnosis of people reacting their experiment.

Summary

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.
Not recommend using the Bonferroni correction. Because this experiment evaluation metrics are correlated. The individual alpha level is 0.05 and if we use Bonferroni correction. The overall alpha level is 0.05 which means the individual alpha level is 0.025. It is too conservative.
If only one metric moved, it should depend on the situation. All the company has its own direction. They can decide to launch the change or not according to company direction.
For the gross metric, we expect this experiment can reduce the number of students who don’t have enough time to finish the class. And the result shows that both statistical significance and practical significance.
For the net conversion, we expect not significantly reducing the number of students to continue past the free trial and eventually complete the course. So this metric should not significance and the statistic also shows the same result.
The discrepancy between effect size and sign test depend on the situation. Sometimes this effect by the weekend or holiday or some huge change on a specific day. Find the reason and make an analysis again.

Recommendation

Due to gross metric is both statistical significance and practical significance. And Retention is statistical significance. Gross also pass the sign test. This experiment shows us adding a screenshot will not reduce too many students to stop the free trial. It has a huge contribute to reducing the gross conversion and contribute a little bit for retention. But nothing to do with net conversion.
The 95% CIs lower boundary is -0.0291 and the upper boundary is -0.0120. The CI does not include zero that is statistical significance and the upper bond is less than d_min which means practical significance. On the other hands, gross is also passed the sign test. The gross conversion decrease substantially. Meet our expectation and recommend to launching the change.
The 95% CIs lower boundary is -0.0116 and the upper boundary is 0.0019. The CI contains 0 as well as this metric not pass the sign test. Meet our expectation and recommend to launching the change. But the most of its confidence interval is in the negative space, meaning that there is the possibility that it went down. And if this occurred, it means that it goes against one of the expected behaviors we need to verify on the experiment, and thus, we should vote against the launch.

Follow-Up Experiment

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.

The follow-up experiment could be if we adding a progress bar for the each of class whether contribute the student to complete the course. The hypothesis is to set a clear progress of student archived that might increase the number of students to finish the whole class.
We can use the number of user-id who complete checkout as invariant metric. Because we only want to measure the student who pass the course.
We use the number of user-id who finished the class divided by the number of user-ids who complete checkout as the evaluation metric.
Unit of the deviation could be number user id who complete the course. Because we want to measure the student who pass the course.