1. Experiment Design

1.1 Metric Choice

> *List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)*
> *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.*

Invariant Metrics:

- **Number of cookies**(*number of unique cookies to view the course overview page*): I choose it as an invariant metric because it shouldn't be changed, and I can use it as unit of diversion. Since it shouldn't be changed, I wouldn't choose it as an evaluation metric.
- **Number of clicks**(*number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger)*): Since it happens before the free trial screener is trigger, amounts of users proceed from course page to the experiment page would be the same between groups (control, experiment). So I choose it as an invariant metric, not evaluation metric.
- **Click-through-probability**(*number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page*): I choose it as invariant metric, since the clicks are occurred before the users see the experiment, and don't depend on our test. Since it shouldn't be changed, I wouldn't choose it as an evaluation metric.

Evaluation Metrics:

- **Gross conversion**(*number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button*): Since it's directly dependent on the effect of the experiment, and it might be changed between groups (control, experiment). I choose it as an evaluation metric, not invariant metric. In detail, users will make a decision of whether to enroll the free courseware while seeing the pop-up message. Which will might be decrease the number of users who enrolling the free courseware.
- **Retention**(*number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout*): Since it's also dependent on the effect of the experiment, and it might be changed between groups (control, experiment) , as mentioned in gross conversion metric. It could be chosen as an evaluation metric, not invariant metric. however, since it will take too long to finish experiment with retention as I will refer in sizing part, I will not use it as an appropriate metric.
- **Net conversion**(*number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button*): Again, it's dependent on the effect of the experiment, and it might be changed between groups (control, experiment). I choose it as an evaluation metric, not invariant metric.

Useless Metrics

- **Number of user-ids**(*number of users who enroll in the free trial*): I will not choose it as an invariant metric. The number of user-ids depends on the experiment, so it can't be used for diversion and might be changed between groups (control, experiment). Also, I will not use it as an evaluation metric. Because it's a raw count. The number of visitors between experiement are likely to be different, which will skew the results. Ratio will be more meaningful than raw number as an evaluation metric.

In order to launch the experiment, we expect the following result:

- The gross conversion will decrease practically significance, which indicate whether the cost will be lower by introducing the screener.
- The net conversion will not decrease statistically significance, which indicate the screener whether or not affect the revenues.

1.2 Measuring Standard Deviation

> *List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)*
> *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.*

- **Gross conversion**: 0.0202
- **Retention**: 0.0549
- **Net conversion**: 0.0156

Since both metrics using number of cookies as denominator, the unit of diversion is equal to unit of analysis. I think the analytical estimate would be comparable to the emperical variability. Meanwhile, the denominator of retention is not equal to unit of diversion, I think the analytical estimate wouldn't be comparable to empirical variability.

1.3 Sizing

1.3.1 Number of Samples vs. Power

> *Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)*

I will not use Bonferroni correction, since metrics in the test has high correlation and the Bonferroni correction will be too conservative.

Specification for not using retention as an evaluation metric any more
After some esitmation, I found the number of pageviews for retention was very large. And it might need over a hundred days for testing. Since it takes too long to finish experiment, I now leave retention out as an evaluation metric.
So let's see the number of pageviews needed for the left two evaluation metrics.

	base conversion rate	dmin	alpha	beta	sample size per group	CTP	pageview per group	total pageview
Gross conversion	0.20625	0.01	0.05	0.2	25835	0.08	322937.5	645875
Net conversion	0.1093125	0.0075	0.05	0.2	27413	0.08	342662.5	685325

Note: when calculating sample size per group, I use tool on site http://www.evanmiller.org/ab-testing/sample-size.html .
From the table, we can see that I need 685,325 pageviews.

1.3.2 Duration vs. Exposure

> *Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. (These should be the answers from the "Choosing Duration and Exposure" quiz.)*
> *Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?*

I divert 0.9 of traffic to this experiment. 685325 / (40000 * 0.9) = 20, So it would need 20 days to run the experiment.
Since I think the experiment will not affect Udacity content, and there is no ethically problematic in this experiment. The experiment doesn't collect sensitive information and will not harm anyone. I think this experiment has no risky for Udacity, and we can divert traffic to experiment with a high rate .

2. Experiment Analysis

2.1 Sanity Checks

> *For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)*
> *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**
  95% confidence interval : [0.4988, 0.5006]
  Observed value : 0.5006
  Result : pass
- **Number of clicks**
  95% confidence interval : [0.4959, 0.5041]
  Observed value : 0.5005
  Result : pass
- **Click-through-probability**
  95% confidence interval : [-0.0013, 0.0013]
  Observed value : -0.00006
  Result : pass

2.2 Result Analysis

2.2.1 Effect Size Tests

> *For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)*

- **Gross conversion**
  95% confidence interval : [-0.0291, -0.0120] statistically significant : Yes practically significant : Yes
- **Net conversion**
  95% confidence interval : [-0.0116, 0.0019] statistically significant : No practically significant : No

2.2.2 Sign Tests

> *For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)*

- **Gross conversion**
  p-value : 0.0026 statistically significant : Yes
- **Net conversion**
  p-value : 0.6776 statistically significant : No

2.2.3 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.*

I didn't use Bonferroni correction, since the metrics in the test has high correlation and the Bonferroni correction will be too conservative.
I completely comprehend the importance to correct if a test is launched and the metrics shows a significant difference, because it's more likely that one of multiple metrics will be falsely positive as the number of metrics increases. However, we would only launch if all evaluation metrics must show a significant change. In that case, there would be no need to use Bonferroni correction. In other words, correction is applied if we are using OR on all metrics, but not if we are testing for AND of all metrics.

2.3 Recommendation

> *Make a recommendation and briefly describe your reasoning.*

The result showned that gross conversion decreased practically and statistically significant under the experimental condition, while net conversion didn't decrease practically and statistically significant, which was expected. But I don't recommend to launch the experiment, since the confidence interval of net conversion includes negative number. It means a probability that the pop-up message can make those who may continue the course after the free period giving up starting free trial. That may decrease the revenue.

3. 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.*

I think follow up experiment can focus on "Retention" (*number of userids to remain enrolled past the 14 day boundary (and thus make at least one payment) divided by number of userids to complete checkout*) by another way.
I suggest a motivation method. It will show a board in thhe browser, which will give students continues update on how much time they have spent on the course (and the time they've planed), and a motivational trigger to keep learning. The trigger is the comparision between the user and others(users in experiment group), like the rate of users you've surpass depend on completion progress, who enrolled almost at the same time with you.

In an experiment the display of such board could be tested. The experiment group would see the board in the browser and the control group not.

- **Hypothesis** : The display of such board would decrease users' frustration and motivate users to seize time to learn.
- **Unit of diversion** : Userids. I choose userids since I divert them after they signed up and gave their credit card information
- **Invariant metric** : Number of userids enrolled into the free trial.

- **Evaluation metric** : Retention rate.

Reference

https://en.wikipedia.org/wiki/A/B_testing
http://www.evanmiller.org/ab-testing/sample-size.html