Experiment Design

Metric Choice

Invariant metrics:

1. Number of cookies (pageviews): Proper randomization should ensure that the correct (here equal) proportion of cookies is assigned to each of the two groups. There is no exposure to the experimental condition before this step (viewing the course overview). This should be invariant not only across the control and experimental groups, but across different platforms and browsers. This is an important sanity check for the experiment, since unequal assignment to the two groups would indicate something seriously wrong with the randomization.
2. Number of clicks: This should be invariant if #1 is invariant. There should be no difference in the user experience prior to clicking the "start now" button, and thus no reason for a significant difference here. Finding a difference in number of clicks would imply that there had been some difference in user experience, or failure of randomization, thus this is also a sanity check.

Evaluation metrics:

1. Gross conversion: This is a primary evaluation metric. The experimental design specifically aims to deter students with insufficient time from enrolling in the course. This is the first part of the experimental hypothesis.
2. Net conversion: The second part of the hypothesis: "without significantly reducing the number of students to continue past the free trial and eventually complete the course" implies that this metric should not be negatively impacted by the experimental condition. The business goal is to have students begin the paid portion of the course (and complete it). The experiment should not deter students who would otherwise progress through the free trial and begin payment.

Unnecessary metrics:

1. Click-through probability: This is the ratio #1/#2 from invariant metrics. If the first two are invariant, this would be also, and not needed as a separate metric.
2. Retention: Decreasing early cancellations (frustrated students) is the goal of the experiment, which means an increase in retention. The means, however, is to deter those with insufficient time from starting in the first place (decrease gross conversion). Retention is net conversion divided by gross conversion, and in this design is a compound measure which does not need to be evaluated separately.

Not useful as a metric:

1. Number of user-ids: By itself, the number of users who enroll in the free trial wouldn't be useful. The absolute number could vary by day, or other condition, and needs to be related to another metric, such as number of cookies viewing the page, before it would be useful.

In order to recommend launching the experiment, I would want to see:

1. A statistically significant decrease in gross conversion at least as large as the practical significance.
2. No decrease in net conversion. The experiment should be sufficiently powered to detect a change at the practical significance level with the given beta (Here 0.2 or a power of 80%). The experiment should not be undertaken if the time and resources needed to achieve this power are not available.

Measuring Standard Deviation

Given a baseline of 5000 page views and the baseline numbers provided, the calculated standard deviation for each of the proposed evaluation metrics would be:

Gross conversion: 0.0202

Net conversion: 0.0156

The unit of diversion for this experiment is a cookie. The unit of analysis for both ot these metrics is a cookie. Since the unit of diversion and unit of analysis are the same for both, the analytic estimate of variability is likely to match the empirical variability.

Sizing

Number of Samples vs. Power

The Bonferroni correction is not indicated for this analysis. There are 2 metrics being analyzed. Both need to meet criterion to recommend launch, and the errors of greater concern are different for each metric. The sizing calculations from the baseline data provided give these estimates of required pageviews: gross conversion - 645875 and net conversion - 685325. The larger of these is the estimate used to calculate duration.

Duration vs. Exposure

Risk assessment:

The risk of harm to users seems low. The only downside is that a student might be inappropriately deterred by the intervention. This seems minor, and unlikely for the type of student seeking independent online study.

There is no additional sensitive or private information being collected in the experimental condition. Enrollment, of course, requires obtaining financial data, but students in the experimental group are actually expected to be less likely to provide this information. There should already be appropriate data security measures in place to protect financial data.

The risk of harm to Udacity's reputation by the presence or absence of a cautionary step before enrollment seems quite low.

Since this is a low risk experiment, I would divert 100% of traffic to the experiment, if possible. Using the larger number above to estimate duration gives 18 days as the time needed to run the experiment.

Experiment Analysis

Sanity Checks

The invariant metrics are the number of pageviews and the number of clicks in each group. The sanity check analysis is as follows:

Pageviews: Expected fraction assigned to control Fc = 0.500, actual = 0.5006. Confidence interval is 0.4988 - 0.5012. Passes.

Clicks: Expected fraction assigned to control Fc = 0.5000, actual = 0.5005. Confidence interval is 0.4959 - 0.5041. Passes.

Result Analysis

Effect Size Tests

Evaluation metrics:

1. Gross conversion: 95% confidence interval of the difference is -0.0291 to -0.0119. This is statistically significant as the confidence interval does not include 0, and practically significant as it does not include the dmin of 0.01.
2. Net conversion: 95% confidence interval of the difference is -.0116 to 0.0019. Not statistically significant as the confidence interval includes zero. Not practically significant in terms of a positive result, as the confidence interval also includes the dmin of -0.0075. Practical significance may have a different meaning in terms of a negative result - see below.

Sign Tests

The sign test for gross conversion gives a P value of 0.0026, which is statistically significant.
The sign test for net conversion gives a P value of 0.6776, which is not statistically significant.

Summary

The Bonferroni correction was not used for the analysis. The Bonferroni correction is used to reduce the possibility of a type 1 error when multiple metrics are evaluated simultaneously. It makes the criterion for rejecting the null for any one metric more stringent than it would otherwise be. This is important when any one positive finding would be actionable. In this experiment, both conditions need to be met before making a recommendation to launch. If we were looking for a positive result in both conditions, with alpha = 0.05, the probability of a false positive in both metrics (assuming independence) would be P=0.0025. This is already quite small and does not need further reduction. In addition the error of greater concern for one of the metrics is a type 2 error, which the Bonferroni correction does not address.

The effect size analysis shows a significant effect for gross conversion, and does not show a significant effect for net conversion. The sign test analysis gives the same results. However, there is a significant part of the confidence interval for net conversion which is beyond the negative practical significance level for this metric. That is, there is a possibility that the true value for net conversion is in fact beyond the practical significance level. The negative practical significance level extends beyond the mean difference for net conversion by -0.0027, divided by

the SE for net conversion of 0.0034, gives a z score for this point on the distribution for net conversion (assuming normal) of -0.79. This would give a probability of about 0.21 that the true value lies beyond the practical significance boundary.

Recommendation

The experiment showed a statistically and practically significant effect on gross conversion, which was one of the conditions for recommending launch. It did not detect a statistically significant effect on net conversion, which was the other condition for recommending to launch. As noted above, however, there is still a nontrivial probability that the true value for the change in net conversion does exceed the practical significance level. This indicates some risk for loss of revenue for Udacity, at least in the short term. On balance it is probably better not to launch this experiment, but this decision should be made by business leaders.

Follow-Up Experiment

This experiment showed that it was possible to deter some students from enrolling by asking them to consider whether they had enough time to devote to the course. It did not detect a definite negative effect on net conversion, but the results were consistent with that outcome with a significant probability. Perhaps an intervention focused on increasing retention, rather than decreasing enrollment, would have a positive effect.
Hypothesis: some students get off to a poor start by not devoting regular time to the course every day.
Experiment: after enrollment, assign by user-id to control and experimental conditions.
Experimental condition: on first access to the course material on any day during the free trial, if the interval since last access is over 24 hours, pop up a reminder saying something like: "Students' likelihood of success is much higher if they can devote regular time to this course every day." If it is too complex to base this on interval since last access, just have it appear once a day regardless.
Metrics: enrollments, retention. Enrollments should be invariant across the groups, as equal numbers of enrolled user-ids should be assigned to each group. The evaluation metric would be retention (fraction of enrolled user-ids to make first payment). The unit of diversion and unit of analysis would be user-id, as this would remain stable after enrollment. The same user could not enroll in the free trial for the same course again, and therefore would not risk being exposed to both conditions.

Resources used in preparing this analysis:
Udacity course materials;
http://www.evanmiller.org/ab-testing/sample-size.html for sample size calculations.