Experiment Design

Udacity courses come in two variants. There is a free of charge version that offers access to the learning material but has no support from coaches. The paid option has coaching support and a verified certificate after a successful final project review.

The paid offering has a 14 days trial period where the student has access to the full material and support from coaches. After that trial period the student is automatically charged unless they cancel first.

During the trial phase resources are bound at Udacity (e.g. the coaches). Therefore it is important to minimize the number of students that are testing the offer but are not changed into paying customers.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual.

If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of 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.

If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course. The number of enrollments will be reduced without significantly reducing the number of students who continue past the free trial.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page. The uniqueness of a cookie is determined per day.

Metric Choice

Invariant metrics

Invariant metrics should stay invariant through the experiment so their distribution should be comparable for the control and experiment groups. Checking the invariant metrics serve the purpose of verifying the sanity of the experiment.

NUMBER OF COOKIES / PAGEVIEWS

The number of unique cookies to view the course overview page should stay invariant through the experiment as the user sees the overview page before experiencing the change made.

NUMBER OF CLICKS

The number of unique cookies to click the "Start free trial" button should also be invariant because the click happens before the new free trial screener is trigger.

CLICK-THROUGH-PROBABILITY
Click-through-probability is the number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. It should also stay invariant through the experiment as the two metrics it is based upon are invariant too.

Evaluation metrics

The hypothesis of this experiment has two parts:

1. The number of students that enroll is reduced by the additional barrier they have to pass. The barrier sets clearer expectations upfront and reduces the number of students that leave the free trial because they don't have enough time.

2. The number of students that have enrolled (after checking their willingness to spend at least 5 hours per week) and stay enrolled past the 14-day boundary is not significantly reduced. The additional barrier has no significant negative effect.

Therefore the following metrics have been chosen for evaluation:

GROSS CONVERSION
Gross conversion is the 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. These students agreed to spend at least 5 hours to the learning. It is expected that this metric is reduced by the experiment. The change should be at least 1% to be practical significant (dmin= 0.01).

RETENTION
Retention is the number of user-ids to that remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. This metric is expected to be higher for the experiment because the additional barrier filtered out some students that are not committed to do the learning or don't have enough time to follow a course. A higher percentage of students that pass the barrier should therefore stay enrolled after 14 days. The additional barrier could lead to better utilization of Udacity's resources (e.g. the coaches) and better user experience. Therefore the experiment could improve the retention metric. The change should at least be 1% to be practical significant (dmin=0.01).

NET CONVERSION
Net conversion is the 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. This would measure the ratio of students that stay enrolled beyond 14 days by the number of students that click before seeing the additional barrier. The barrier should not have a negative effect, so no significant change is expected. To be practical significant, the change has to be higher than 0.75% (dmin=0.0075).

Unused Metrics

NUMBER OF USER-IDS
This metric counts the number of users who enroll in the free trial. It would be interesting to observe during the experiment, but as the unit of diversion is a cookie and the diversion between experiment and control group is done on cookies before students enroll an even distribution is not expected. Therefore it is not ideal as an evaluation metric but could be used to evaluate how many students stay enrolled beyond the 14-day trial period. I have not chosen to use this metric for evaluation because it is not normalized. It is also not suitable as an invariant metric because it is calculated after the diversion of groups and is not expected to stay invariant.

Measuring Standard Deviation

The analytical estimate of the standard deviation for the evaluation metrics is shown in the following table.

Evaluation Metric	Standard Deviation
Gross Conversion	0.0202
Retention	0.0549
Net Conversion	0.0156

If the unit of diversion is equal to the unit of analytics the analytical estimation of the standard deviation is typically close to the empirically calculated standard deviation. This is the case for the gross conversion and net conversion metrics. Both are based on cookies which is also the unit of diversion for this experiment. The retention metric is based on user-ids which is different from the unit of diversion. Therefore the empirical standard deviation might differ from the above number and should be calculated if retention is used to judge the experiment.

Sizing

Number of Samples vs. Power

To calculate the number of samples to achieve statistical power I used the website http://www.evanmiller.org/ab-testing/sample-size.html. For all calculations an alpha level of 5% and beta of 20% is used. The resulting sample sizes have to be doubled because two sample sets are needed for the control and experiment group. In order to calculate the needed page views (or cookies) for the experiment the relationships in the baseline data between the metrics and pageviews is used. The following table shows the results for the evaluation metrics.

	dmin	baseline	sample size	needed pageviews
Gross Conversion	1 %	20.625 %	25835	645876
Retention	1 %	53 %	39115	4741212
Net Conversion	0.75 %	10.93125 %	27413	685326

The Bonferroni correction would be needed to reduce type I errors or false positives if a decision is made on any (or not all) of the observed metrics. As both remaining metrics are required for a decision a Bonferroni correction is not needed.

Duration vs. Exposure

The retention metric would require more than 118 days to achieve statistical power even if 100% of the traffic would be diverted to the experiment. Spending more than 118 days for a single experiment would imply a high risk for the business because it is unclear for month what the results might be and other experiments are also not possible. Therefore it is obvious that retention is not suitable as an evaluation metric and just the gross conversion and net conversion metrics are taken into consideration.

Although the change for this experiment is de facto a barrier that students have to pass before they become paying customers the inherent effect is expected to be low. Therefore 100% of the traffic could be diverted to the experiment. This would result in a duration of 18 days to achieve sufficient

statistical power. Diverting just 50% of the traffic to this experiment would extend the duration to more than 37 days. This could still be acceptable and would make it possible to also run a parallel experiment.

Experiment Analysis

Sanity Checks

For all the invariant metrics the observed values from the control and experiment group show that they are in fact invariant and pass the sanity checks. A 95% confidence interval was used for the sanity checks. The following table shows the details.

	Control	Experiment	Total	Expected	Observed	CIlower	Clupper	Pass
pageviews	345543	344660	690203	0.5	0.5006	0.4988	0.5012	True
clicks	28378	28325	56703	0.5	0.5005	0.4959	0.5041	True
click through				0.0822	0.0821	0.0813	0.0831	True

Result Analysis

Effect Size Tests

To calculate the effect of this experiment only a part of the data could be used, because only 14 days after the start of the experiment data for the evaluation metrics is available. It should be noted that only 423525 pageviews were detected during the experiment. This is far below the number calculated to get statistical power.

To analyze the differences between the control and experiment group a 95% confidence interval was used for the evaluation metrics. Based on the observed data, the change for the gross conversion metric is as expected. The gross conversion was reduced and the reduction is statistically and practically significant. For the net conversion metric the observations do not show any statistically or practically significant change. The following table shows the details:

	Dmin	Diff	CIlower	Clupper	StatSig	PractSig
gross conversion	0.0100	-0.020555	-0.029123	-0.011986	True	True
net conversion	0.0075	-0.004874	-0.011605	0.001857	False	False

Sign Tests

The day by day sign test shows 19 negative days for gross conversion but only 13 positive days for net conversion. Using an alpha of 5% the sign test shows statistically significance for cross conversion only. The following table shows the details:

	p value	StatSig
gross conversion	0.0026	True

	p value	StatSig
net conversion	0.6776	False

Summary

As already mentioned above the Bonferroni correction is used to reduce the probability of type I errors (false positives) when multiple metrics are used to test a hypothesis. It is necessary if just any metric is used to test the hypothesis. For this test, a single metric is not sufficient to test the hypothesis. All metrics have to be statistically and practically significant to launch the change. Therefore a Bonferroni correction is not appropriate.

The sanity checks show that the experiment is not influenced by other effects not under control. The evaluation metrics could not proof the hypothesis. Just the change of the gross conversion metric is statistically and practically significant. The net conversion metric does not show any significant change for the experiment. The sign tests support these findings.

The experiment run for 37 days and from the baseline data 50% of the traffic was diverted to it. Therefore the results have the necessary statistical power.

Recommendation

The experiment shows a significant reduction of the gross conversion. So there is a reduction of students enrolling a course. The barrier works. There is also no significant reduction for net conversion, the number of students that keep enrolled beyond the 14 days trial period. But the 95% confidence interval for net conversion does include the negative of the practical significance boundary for the metric. This means, it is possible that the number of students that keep enrolled beyond the 14 days trial period is reduced by an amount the business cares about.

Overall, the experiment shows statistically and practically significant reduction of the gross conversion. This supports part 1 of the hypothesis. But as a reduction of the net conversion is possible, part 2 of the hypothesis is not supported by the evaluation. It is therefore not advisable to launch the change.

As noted above, the duration of the experiment is not sufficient to achieve statistical power. So if feasible the experiment should continue to get the needed statistical power and evaluate the metrics again. If it is not feasible to spend more time for the experiment other experiments should be considered.

Follow-Up Experiment

The intention of the above experiment is to free up resources and improve the user experience. The resources are freed up by a barrier that reduces the number of students enrolling. The user experience should improve because students are aware of the effort that is expected and the better support they have from the coaches. But the overall goal is to improve the number of paying customers. The approach to achieve this is passive because there are no additional actions by Udacity to improve the user experience.

I would recommend an experiment that is more active in improving the user experience. An idea I have is to have online chat sessions (or hangouts if that is possible for many people following) during the free trial period that help students ramping up. The chat sessions could be about technical topics (e.g. working with the platform), about the content (discussing prereqs, overview of course, etc.), or about alternative courses (e.g. lower/higher level).

The unit of diversion would be the user-id. After enrolling into a course the students are divided into an experiment group that has this additional chat sessions and a control group that hasn't. The hypothesis of this experiment is that students who have the chance to get help from the chats will more likely stay enrolled after the 14-day trial period. This could be because they really got help or they expect to get help in the future when needed.

The evaluation metrics for this experiment would be the retention rate, the number of user-ids that remain enrolled after the 14-day trial period divided by the number of user-ids that completed checkout and enrolled in a course. It is expected that the retention rate rises by the experiment. As the effort to run the experiment is high, a higher change is needed to justify launching this experiment. 1% is definitely not enough but I have no experience to say if 5% or 10% is a reasonable value for dmin.

The pageviews/cookies, clicks, click-through-probability, user-id, and gross conversion metrics are all invariant because they are measured independently of the diversion into control and experiment groups. The net conversion metric is not suitable to evaluate the test because the analytic basis is different from the unit of diversion.

Counting the number of user-ids that attended a chat could be an additional metric. This metric is not suitable for evaluation of the experiment because there are no values in the control group. Nevertheless it could be interesting to see if the additional offering of chats is accepted at all.