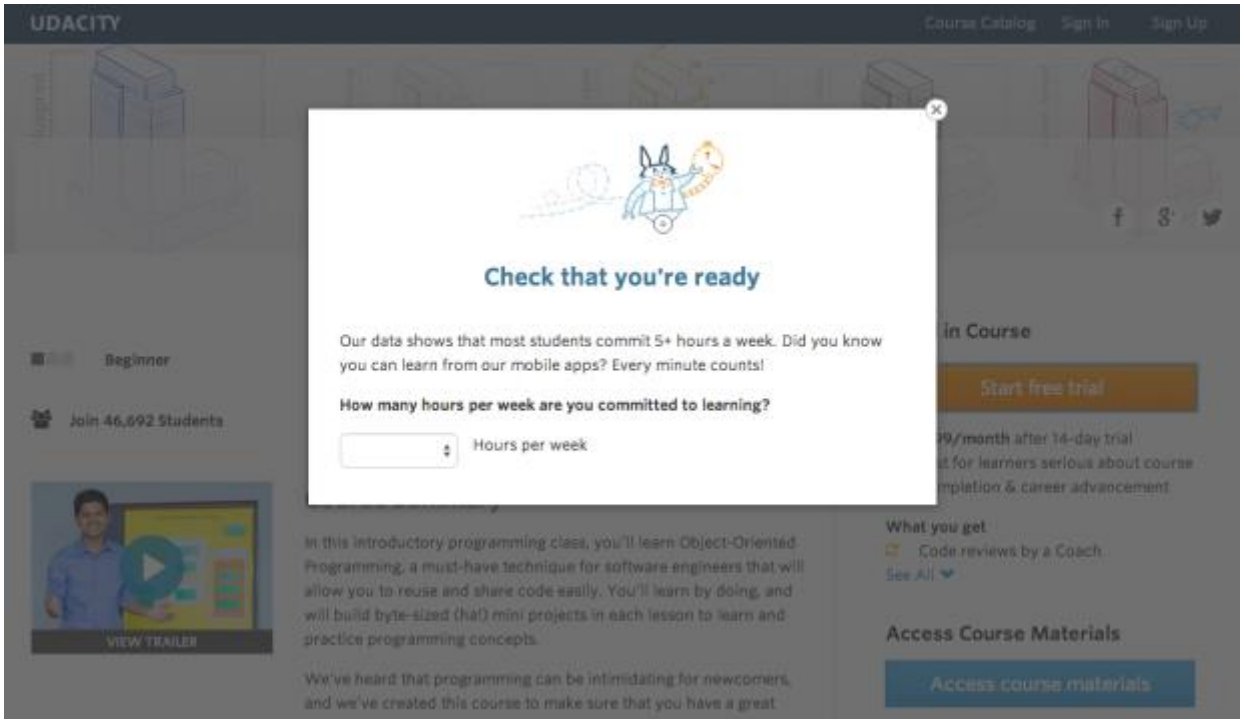
Design an A/B Test

1. Experiment Overview:

Udacity courses have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", unless they cancel within 14 days, they will automatically be charged for the course paid version. In this 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 go through as usual. Otherwise, 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. The following screenshot shows what the experiment looks like:



The hypothesis is that this will reduce the number of frustrated 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. Then Udacity could improve the overall student experience as well as the coaches' capacity to support students who are likely to complete the course.

2. Experiment Design

The unit of diversion is a unique cookie (the same cookie visiting on a different day would be counted twice). Although if the student enrolls in the free trial, they are tracked by user-id from that point forward, where user-ids are automatically unique since the site does not allow the same user-id to enroll twice.

Metrics Choice

Invariant Metrics

- Number of cookies: That is, the number of unique cookies to view the course overview page. This is a good invariant metric since it is randomly, thus evenly distributed to the control and experiment group.

- Number of clicks: That is, the number of unique cookies to click the "Start free trial" button. Since this occurs before the free trial message, again we have a random distribution across the control and experiment group.

Evaluation Metrics

- Gross conversion: That is, the number of user-ids to enroll in the free trial, divided by the number of unique cookies to click the "Start free trial" button. This metric will be used as an evaluation metric to indicate the extent, that the experiment discouraged students to enroll for a free trial to the paid course version. And even though we expect this metric to decrease as a result of the experimental feature, in order to run the experiment, we need to ensure that this decrease will be limited.

- Net conversion: That is, the number of user-ids to remain enrolled past the 14-day boundary, divided by the number of unique cookies to click the "Start free trial" button. This is a straightforward metric, which will indicate if and to what extent, the experiment has reduced the number of students, who went through the trial period to the first payment. This is a critical factor for determining whether this experiment will be actually launched, and in order to do so, Net conversion must not reduce.

Not used Metrics

- Number of user-ids: That is, number of users who enroll in the free trial. It isn't suitable as an invariant metric, due to the fact that it is not evenly distributed to the control and experiment group. And even though it could be used as an evaluation metric, I believe the ones chosen will provide us with more insights on the experiment's objectives.

- Click-through-probability: That is, number of unique cookies to click the "Start free trial" button, divided by the number of unique cookies to view the course overview page. This metric is a fraction of the two invariant metrics already chosen and therefore, even though it would make a good invariant metric, it wouldn't really add additional information to the experiment.

- Retention: That is, number of user-ids to remain enrolled past the 14-day boundary, divided by the number of user-ids to complete checkout. This a metric consist as well, from fractions of the two metrics chosen as evaluation metrics.

Measuring Standard Deviation

- Unique cookies to click "Start free trial" per day: 400
- Enrollments per day: 82.5
- Probability of enrolling, given click: $\frac{82.5}{400} = 0.206$
- Probability of payment, given enroll: 0.53
- Probability of payment, given click: $0.206 * 0.53 = 0.109$

Gross conversion σ : $\sqrt{0.206 * \frac{1-0.206}{400}} = 0.0202$

Net conversion σ : $\sqrt{0.109 * \frac{1-0.109}{400}} = 0.0156$

Since the unit of analysis (number of unique cookies to click the "Start free trial" button) is for both the same as the unit of diversion (a unique cookie) we expect the analytical estimates to be comparable to the empirical variability. Thus, even if we had the time we wouldn't need to also go for an empirical estimate.

Sizing

Number of Samples vs. Power

- α : 5%
- β : 20%
- Gross Baseline conversion rate: 20.625
- Gross Minimum detectable effect: 1%
- Net Baseline conversion rate: 10.931
- Net Minimum detectable effect: 0.75%

During our analysis, we will not make use of the Bonferroni correction, since we have only chosen two evaluation metrics. In order to calculate the sample size (minimum page views) to power our experiment we used "Evan Millers' A/B Testing Sample Size Calculator" with the following results:

Evaluation Metric	Page Views Required
Gross conversion	645,875
Net conversion	685,325

Duration vs. Exposure

The number of days the experiment should run is calculated as follows:

- Unique cookies to view page per day (traffic): 40,000
- Minimum total page views required for the experiment: 685,325
- Minimum days: $\frac{685{,}325}{40{,}000} = 17.13$

This experiment definitely won't pose more than the minimum risk for the student participants, since it doesn't involve any sensitive data. Informing the students who enroll for a free trial, on the necessary time commitment, is a pretty harmless change. In addition, the experiment's short duration ensures minimum risk for Udacity organisation too, allowing for a safe diversion for up to 100% of the traffic. Thus we can run the experiment for exactly 3 weeks (21 days instead of 17), in order to assure an even ratio of weekends and weekdays. In this case, the fraction of traffic is calculated as follows:

- Unique Cookies per day: $\frac{685{,}325}{21} = 32{,}635$
- Fraction of traffic diverted to the experiment for 21 days: $\frac{32{,}635}{40{,}000} = 0.816$

3. Experiment Analysis

Sanity Checks

For both our invariant metrics, below are the 95% confidence intervals for the values we expect to observe, the actual observed value, and sanity test pass/fail determination.

Invariant Metric	Number of Cookies	Number of Clicks
Control group	345,543	28,378
Exp. group	344,660	28,325
Total	690,203	56,703
Probability	0.5	0.5
Standard Error	$\sqrt{0.5 * \frac{1-0.5}{345{,}543+344{,}660}} = 0.0006$	$\sqrt{0.5 * \frac{1-0.5}{28{,}378+28{,}325}} = 0.0021$
Confidence Interval	0.95	0.95
Z score	1.96	1.96
m	$0.0006 * 1{,}96 = 0.0012$	$0.0021 * 1.96 = 0.0041$
p-m	0.4988	0.4959
p+m	0.5012	0.5041
Observed Value	$\frac{345{,}543}{690.203} = 0.5006$	$\frac{28{,}378}{56{,}703} = 0.5005$
Sanity Check	**Pass**	**Pass**

Result Analysis

Effect Size Tests

A metric is statistically significant if the confidence interval does not include 0 (we can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (we can be confident there is a change that actually matters to the organisation.)

Evaluation Metric	Gross Conversion		Net Conversion	
Groups	Control group	Experiment group	Control group	Experiment group
Unique clicks	17,293	17,260	17,293	17,260
Enrollments	3,785	3,423		
Payments			2,033	1,945
Ppool	$\frac{3{,}785+3{,}423}{17{,}293+17{,}260} = 0.2086$		$\frac{2{,}033+1{,}945}{17{,}293+17{,}260} = 0.1151$	

SEpool	$\sqrt{0.2086 * (1 - 0.2086) * \left(\frac{1}{17,293} + \frac{1}{17,260}\right)} = 0.0044$	$\sqrt{0.1151 * (1 - 0.1151) * \left(\frac{1}{17,293} + \frac{1}{17,260}\right)} = 0.0034$
dmin	0.01	0.0075
$\hat{d}$	$\left(\frac{3,423}{17,260}\right) - \left(\frac{3,785}{17,293}\right) = -0.0205$	$\left(\frac{1,945}{17,260}\right) - \left(\frac{2,033}{17,293}\right) = -0.0049$
Confidence Interval	95%	95%
Z score	1.96	1.96
m	$0.0044 * 1,96 = 0.0086$	$0.0034 * 1,96 = 0.0067$
$\hat{d}$- m	-0.0291	-0.0116
$\hat{d}$+ m	-0.0120	0.0019
Statistically Significant	YES	NO
Practically Significant	YES	NO

Sign Tests

After entering the number of successes observed, the number of experiment trials as well as the "success" probability we got the following results for the two tailed p-values. For Gross conversion, since it is less than the given alpha level (α=0.05), it statistically significant, on the other hand, for Net conversion it is definitely not.

Evaluation Metrics	Gross Conversion	Net Conversion
α	0.05	0.05
Successes	4	10
Trials	23	23
Probability	0.5	0,5
Two-tailed p-value	$0.0026 < \alpha$	$0.6776 > \alpha$
Statistically significant	YES	NO

Summary

In A/B experiments the more metrics you are measuring, the more likely you are to get at least one single false positive. In order to tackle this false positives, we can use Bonferroni correction, which adjusts the significance level at α/m, for testing each individual hypothesis. In our case in order to launch the experiment, we expect null hypotheses to be rejected for both gross and net conversion, therefore we will not make use of Bonferroni correction.

Moreover, there are no discrepancies observed between the effect size hypothesis tests and the sign tests, since in both the effect of the experiment in Gross conversion metric was evaluated as statistically significant, whereas in Net conversion as statistically not significant.

Recommendation

In this experiment, we have noticed a statistically significant decrease in the Gross conversion evaluation metric, which indicates that during our test fewer students enrolled in the free trial (hypothesis test). This has been one of the objectives of our experiment, to reduce the number of the frustrated students and let Udacity coaches work more efficiently and thus improving the overall student experience.

On the other hand, the results obtained for our second metric - Net conversion, where neither statistically not practically significant, the confidence interval does include the negative of the practical significance boundary. Under these circumstances, we decide against launching this experiment, since it is not impossible that this number that could go down to a level that could jeopardize Udacity's revenues.

4. Follow-Up Experiment

Overview

To reduce the number of frustrated students who cancel early in the course, we could try to implement an alternative experiment with the following change:

After the student enrolls in the free trial a message would appear indicating that Udacity courses usually require a minimum 5-hour average per week, for successful completion and present the students with the Udacity's clock feature, built to help them engage in their cause of devoting the necessary time to the course.

The Udacity clock is an engaging informative clock, measuring the students' studying time, as well as inform them of the hours left for them to reach the minimum 5-hour goal per week. It comes in the form of a simple app within the student's Udacity profile and also available for downloading to a smartphone. The students are instructed to run the clock each time they begin a studying session for a Udacity course and pause it when the session is over. The clock works as an engaging tool that makes the students want to reach this short-term goal and thus making them achieve a more concentrative and effective studying.

Hypothesis

The hypothesis is that this change will reduce the number of frustrated 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. Then Udacity could improve the overall student experience as well as the coaches' capacity to support students who are likely to complete the course.

Unit of diversion

Since students have already enrolled in the free trial, they are tracked by user-id, which is automatically unique since the site does not allow the same user-id to enroll twice.

Invariant Metrics

- **Number of user-ids:** This is our Unit of diversion, which counts the number of users who enroll in the free trial. This is a good invariant metric since it is randomly, thus evenly distributed to control and experiment group.

Evaluation Metrics

- **Retention:** That is, the number of user-ids to remain enrolled past the 14-day boundary and make their first payment, divided by the number of user-ids who enrolled in the free trial. This metric will be used as an evaluation metric to indicate if and to what extent, the experiment encouraged students to continue to the paid version.