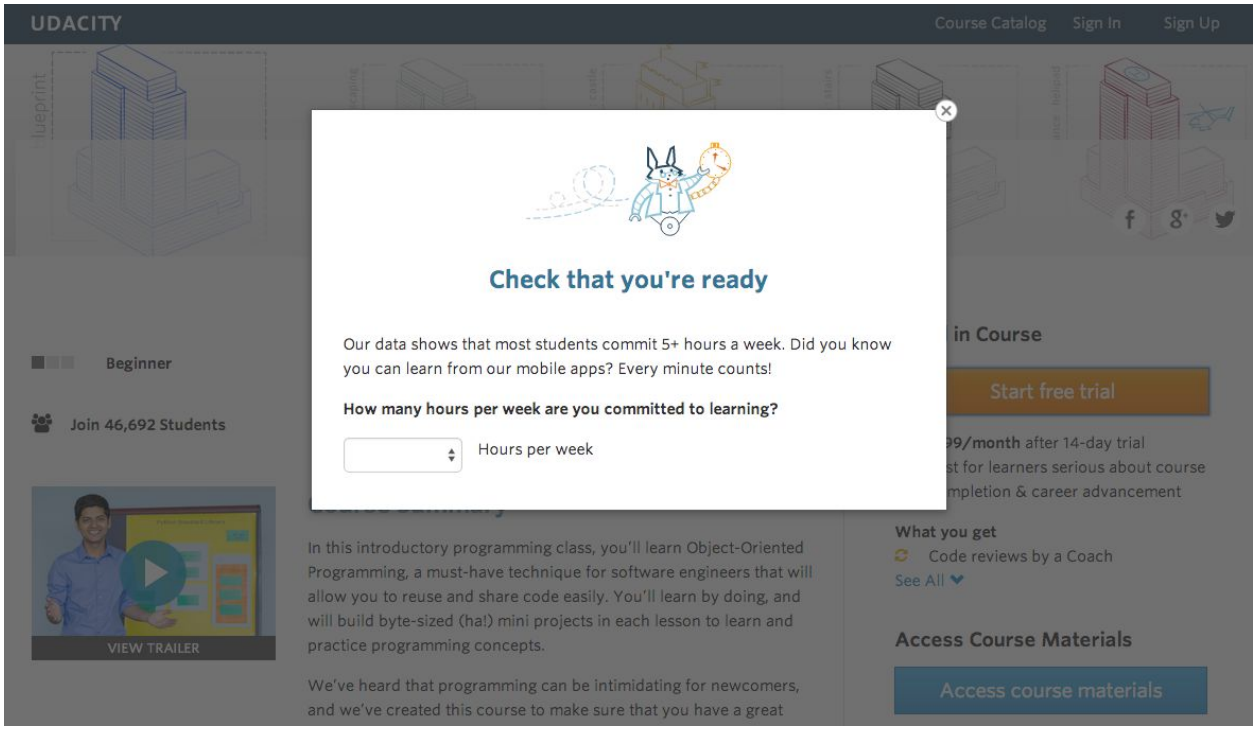
P7: Design an A/B Test

By David Venturi

Hypothesis

Udacity would like to test out the following screen that is activated upon clicking the “Start free trial” button for a course:



The hypothesis is that the free trial screen will reduce the number of frustrated students who leave 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. Udacity could then improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

Experiment Design

Metric Choice

The unit of diversion is a unique cookie (where uniqueness is determined by day), but users are also tracked by user ID if they enroll in a free trial. Invariant metrics are metrics that shouldn’t be affected by the experimental change.

Invariant Metrics

- Number of cookies: This is a good population-sizing metric since the number of cookies assigned to the control and experiment groups should be random and approximately equal.

- Number of clicks: Since clicks are rendered before the free trial screen is triggered, number of clicks shouldn't be affected by the experiment.

- Click-through-probability: Since number of clicks and number of cookies are invariant, and click-through-probability is defined as the number of unique cookies to click the "Start free trial" button divided by the number of cookies to view the course homepage, click-through-probability should be invariant as well.

Evaluation Metrics

- Gross conversion (d_{min} = 0.01): We expect the numerator of gross conversion (number of user IDs to enroll in the free trial) to decrease, but the denominator (number of cookies to click "Start free trial") to remain constant. Thus, we expect gross conversion to decrease. It was chosen as an evaluation metric because it reflects the potential reduction of frustrated students who leave the free trial because of time constraints and the improvement of coach capacity to help enrolled students.

- Net conversion (d_{min} = 0.0075): We hope that the numerator of net conversion (number of user IDs to make a payment) doesn't decrease by much, and we expect the denominator (number of cookies to click "Start free trial") to remain constant. Thus, we hope/expect net conversion to remain constant or decrease a little. It was chosen as an evaluation metric because it reflects the potential preservation of paying students who complete the course.

Unused Metrics

- Number of user IDs: The number of users IDs should be lower in the experimental group than in the control group, which makes user IDs a poor invariant. The count of user IDs could be used as an evaluation metric, but the ratio metrics that have user IDs as their numerator are probably better options.

- Retention (d_{min} = 0.01): We hope that the numerator of retention (number of user IDs to make a payment) doesn't decrease by much, and we expect the denominator (number of user IDs to enroll in the free trial) to decrease. Thus, we expect retention to increase. It wasn't chosen as an evaluation metric, however. While it can reflect the potential preservation of paying students who complete the course (numerator), as well as the potential reduction of frustrated students who leave the free trial because of time constraints, we have already captured both in the via gross conversion and net conversion. With both the numerator and denominator expected to move, identifying the

individual effect of each is trickier, as well. Plus, the unit of diversion (a cookie) is different than the unit of analysis (the denominator of retention, which is user IDs who enroll in a free trial). Variability is often higher when this is the case, which would increase our experiment's size and/or lengthen its duration.

Results Required for Launch

We will need to observe the desired effect for both metrics in order to launch the experiment. These goals are distinct and we can't launch when we only observe one. Again, we need to see:

- A practically significant decrease in gross conversion (d_{min} = 0.01)
- That net conversion doesn't decrease below the practical significance boundary (d_{min} = 0.0075)

See the "Hypothesis" section for a reminder on why these metrics are important.

Measuring Standard Deviation

Both evaluation metrics are probabilities. The standard deviation can be found by taking the square root of the following variance equation, where p is the probability and N is the sample size of the unit of analysis (number of cookies who click on "Start free trial") given a sample size of 5,000 cookies visiting the course overview page.

Calculating variability

type of metric	distribution	estimated variance
probability	binomial (normal)	$\frac{p(1-p)}{N}$

p is found in this table of baseline values. N can also be found there by dividing the number of clicks on "Start free trial" by the same divisor (8) that scales the number of cookies in the baseline table (40,000) to the aforementioned sample size of 5,000.

Gross Conversion

Probability of enrolling, given click:

- p = 0.20625
- N = 3,200 / 8 = 400
- Standard deviation = 0.0202

Net Conversion

Probability of payment, given click:

- p = 0.1093125
- N = 3,200 / 8 = 400
- Standard deviation = 0.0156

The unit of analysis (i.e. the denominator of both evaluation metrics) is the number of unique cookies to click the "Start free trial" button. Since the unit of diversion (a cookie) is the same as the unit of analysis for both metrics, the analytical estimates of variability will likely be close to the empirical estimates. We shouldn't need to do empirical estimates, even if we have the time.

Sizing

Number of Samples vs. Power

Bonferroni Correction

The Bonferroni correction will not be used during the analysis phase. The purpose of the Bonferroni correction is to adjust for experiments with multiple metrics where we only require a few metrics to be significant to launch. In these cases, the probability of obtaining at least one false positive increases (the family-wise error rate) as the number of metrics increases, so we make it less likely that we detect false positives by using the Bonferroni correction. As Udacity Coach Sheng Kung phrases it: "the fewer metrics that you require to be significant to make a decision, and the more independent these metrics are, the stricter you need to be with your significance level to constrain your overall error rate."

The overarching concept that we want to keep in mind is that we want to control our overall probability of making an error in our conclusions. Because we will only launch if all of our evaluation metrics are practically significant, we are already being sufficiently conservative and don't need to use the Bonferroni correction. Since gross conversion and net conversion are dependent on each other, and we are testing to see if the underlying phenomena involving both change, requiring that both be significant (being conservative) and also using the Bonferroni correction (being extra conservative) makes it too difficult to detect true positives.

Number of Pageviews

Using Evan Miller's Sample Size Calculator, the following inputs were used for both evaluation metrics:

- Baseline conversion rate: 20.625% for gross conversion and 10.93125% for net conversion, as per the table of baseline values
- Minimum detectable effect (absolute): 1% for gross conversion and 0.75% for net conversion, as per the project instructions

- Statistical power (1-β): 80%
- Significance level (α): 5%

For both evaluation metrics, the sample sizes required to power our experiment are as follows:

Evaluation Metric	**Unit of analysis sample size per branch**	**Pageview sample size per branch**	**Pageview sample size for overall experiment**
Gross conversion	25,835 cookies who click “Start free trial”	322,937.5 pageviews (25,835 * 12)	645,875 pageviews (322,937.5 * 2)
Net conversion	27,413 cookies who click “Start free trial”	342,662.5 pageviews (27,413 * 12)	685,325 pageviews (342,662.5 * 2)

The larger sample size of 685,325 pageviews is required to sufficiently power the experiment for both evaluation metrics.

Duration vs. Exposure

I chose to divert 81.6% percent of traffic to this experiment. Given that we require 685,325 pageviews and Udacity gets 40,000 page views per day, we require 21 days to run the experiment.

21 days was chosen as the duration because I understood that as the maximum allowable duration according to the project instructions. They state, “If the [duration] is longer than a few weeks, then this is unreasonably long.” The word “few” is often understood as corresponding to the number three.

The experiment could be run for 18 days at 100% traffic, however I chose to extend it to 21 days being aware of the potential negative effects of diverting 100% traffic. These include:

- Data collected over a shorter period of time is at risk of being influenced by specific events (e.g. weekends and holidays). Running the experiment for longer increases our sample size of days and reduces our exposure risk to abnormal days.
- Weekend traffic is often different than weekday traffic. 21 days is an even three weeks, whereas 18 would have lower proportions of weekdays or weekends, depending on the start date.
- There’s a chance there’s a bug in the new feature. Rolling out to a lower percentage of traffic reduces bug exposure assuming we catch it at the same time as rolling out to 100%.

The experiment isn't very risky, in my opinion. Enrolling in a course likely isn't life or death, like changing the settings in a blood sugar monitoring app might be. Also, the addition of the feature also won't affect the user experience once a user decides to enroll, even if the feature breaks, like the change of a database and it subsequently breaking would. The only risk is losing revenue via lost paying customers, but if we are committed to the current design of our experiment, we need to let it play out until we achieve our required number of pageviews (685,325). Limiting exposure doesn't change this fact, it only extends the experiment's duration.

Experiment Analysis

Sanity Checks

For each invariant metric, below are the 95% confidence intervals for the values we expect to observe, the actual observed value, and sanity test pass/fail determination. The values for number of cookies are number of cookies in the control group divided by the number of cookies in both groups, where the expected fraction is 0.5. The same is true for number of clicks. The values for click-through-probability are the difference between the control and the experiment click-through-probabilities, where the expected difference is 0.

Invariant Metric	Lower Bound	Upper Bound	Observed Value	Passes?
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through-probability	-0.0012	0.0013	0.0001	Yes

Result Analysis

Effect Size Tests

For each evaluation metric, below are the 95% confidence intervals around the difference between the experiment and control groups. Statistical and practical significance determinations are made, as well.

Evaluation Metric	Lower Bound	Upper Bound	Statistically Significant? (α = 0.05)	Practical Significance Boundary	Practically Significant?
Gross conversion	-0.0291	-0.0120	Yes	0.01	Yes
Net conversion	-0.0116	0.0019	No	0.0075	No

Sign Tests

For each evaluation metric, the p-value for a sign test is reported alongside a statistical significance determination.

Evaluation Metric	Sign Test p-value	Statistically Significant? (α = 0.05)
Gross conversion	0.0026	Yes
Net conversion	0.6776	No

Summary

The Bonferroni correction was not used. Because we required the effects of both metrics to be practically significant to make a launch decision, we did not need to be more conservative and further decrease our false positive risk. This decrease would have come at the expense of our ability to detect a true effect. Please consult the “Bonferroni Correction” section in the “Sizing” section of this report for a more detailed explanation.

There are no discrepancies between the effect size hypothesis tests and the sign tests. Both declare the difference between the control group and the experiment group for gross conversion as statistically significant, but not for net conversion.

Recommendation

I recommend that the free trial screen is not launched and that we dig deeper. The good news: the confidence interval for gross conversion was entirely below the -1% boundary, which would benefit Udacity by freeing up coach capacity and their ability to improve student experience. The lower bound of the confidence interval for net conversion, however, was below the -0.75% boundary. This potential lack of preservation of paying students would not be good for Udacity’s revenues.

Because a decrease in net conversion below -0.75% is not acceptable for the business, we should dig deeper in this area. The lower bound of the confidence interval for net conversion is -0.0116, which is not that far off from -0.0075. Perhaps we can run the experiment again on a different set of days. Halloween, Canadian Thanksgiving, and the lead up to Christmas could cause an abnormal drop in payments as for some people this is an expensive time of year. We could also run the experiment past our sized number of pageviews to tighten our confidence interval, though we should be aware of increasing our false positive rate.

Follow-Up Experiment

Here is a potential follow-up experiment to reduce the number of frustrated students who cancel early in the course.

Change

Send an email to everyone who enrols in the free trial four days after they enrol in the free trial, which is just over halfway through Udacity's one week free trial period. This email will contain information similar to the free trial screener, where the importance of dedicating 5+ hours per week is emphasized.

Hypothesis

Since this email will be sent five day after enrolling rather than prior to enrolling at all, students will be less likely to quit the course upon hearing the required time commitment. I am also assuming that a cancellation within the free trial period does not count as a frustrated student.

This change could help retain students who did not enroll in the course in the free trial screen experiment because they were intimidated by the time commitment. These students might be able to dedicate less time to be successful, or they might be able to find time in their schedule if they enjoy the course. After trying the course, they might be more likely to stick past the free trial period.

Unit of Diversion

If a student enrolls in the free trial, they can only be reliably tracked by user ID from that point forward. Since the unit of diversion should match up with how we identify users at the point where the dividing mechanism (email or no email) will be implemented, the unit of diversion should be user ID in this case.

Metrics

To measure the reduction of the number of frustrated students who cancel early in the course, we require a modified metric. Since the email is sent post-enrolment, we can't use gross conversion like we did in the free trial screen experiment. The numerator of this modified metric could be number of user IDs to cancel during the free trial period after receiving the email. The denominator (the unit of analysis) would be best fit to match the unit of diversion, which in the follow-up experiment is user ID that enrols in a free trial.

The population sizing invariant metric would be user ID to enroll in a free trial, which would be randomized between our experiment's two branches. Number of clicks and click-through-probability would also be invariant, for the same reasons they are in the free trial screen experiment. Number of cookies would also be invariant since user ID to enrol in a free

trial can correspond to multiple cookies, and users likely don’t a systematic exhibit cookie clearing trend that would spoil the randomization of the user ID diversion.