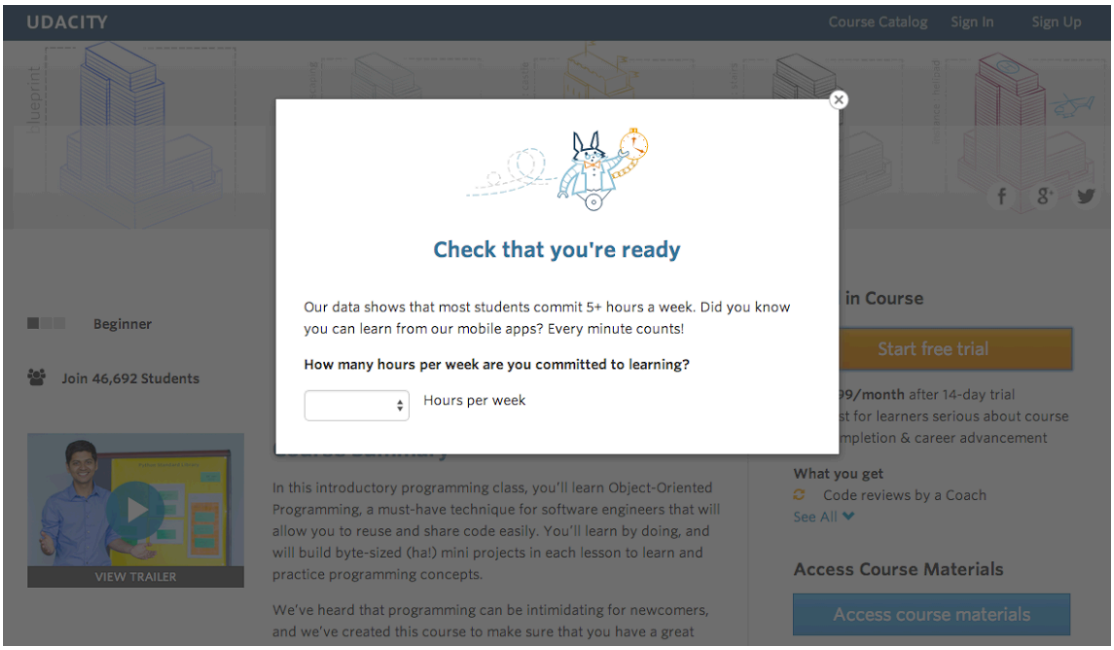
Design an A/B test

Juho Salminen

Experiment design

On the Udacity course homepages students can click a “start free trial” button, after which they are asked to provide their credit card information. Two weeks later, if the students have not cancelled their subscription, their credit card is automatically charged. This experiment aims at finding out whether setting clearer expectations for the enrolling students on the course workload would reduce the number of students quitting the free trial before the first payment. Two conditions are explored. In control condition the site works as described above. In experiment condition after clicking the “start free trial” button students are first asked how much they have time to devout to the course per week, and if their answer is less than 5 hours, they are instructed that the course usually takes more effort than that and they suggested to try the free version. The figure below depicts the experiment site.



The hypothesis is that in the experimental condition the number of frustrated students quitting the trial is reduced without significantly reducing the number of students who continue past the free trial.

Metric choice

The following metrics are collected on the Udacity site. Differences in metrics considered significant in practice are in parentheses.

- Number of cookies: number of unique cookies to view the course overview page. (d_{min}=3000)

- Number of user-ids: number of users who enroll in the free trial. (d_{min}=50)
- Number of clicks: number of unique cookies to click the "start free trial" button. (d_{min}=240)
- 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.(d_{min}=0.01)
- 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. (d_{min}= 0.01)
- Retention: number of user-ids that remain enrolled at least 14 days and as a result make at least one payment divided by number of user-ids to complete checkout. (d_{min}=0.01)
- Net conversion: number of user-ids that remain enrolled at least 14 days and make at least one payment divided by the number of unique cookies to click the "Start free trial" button. (d_{min}= 0.0075)

Table below explains the use of each available metric as an invariant or evaluation metric.

Metric	Invariant metric	Evaluation metric
Number of cookies	**Yes**. As an initial unit of diversion, this metric is used to check that same number of users enter the experiment and control conditions.	**No**. Invariant metric.
Number of user-ids	**No**. This metric is not relevant for diversion.	**No**. Absolute number of user-ids enrolling the free trial depends on the number of users clicking the “start free trial” button in addition to the experiment condition.
Number of clicks	**Yes**. Used to check that the same number of users proceed from course page to the experiment page in the experiment and control conditions.	**No**. Invariant metric.
Click-through-probability	**Yes**. Ensures that the students are similar in both groups in that they continue to the free trial with similar probability. Normalizes the differences between group sizes.	**No**. Invariant metric.
Gross conversion	**No**. Evaluation metric.	**Yes**. Experiment might change this metric, as some of the students are suggested to opt for a free version instead.
Retention	**No**. More suitable as an evaluation metric.	**No**. Retention is not relevant for the goals of the experiment.
Net conversion	**No**. Evaluation metric	**Yes**. Experiment might change this metric by directing some of the students to the free version.

Measuring standard deviation

Baseline values for the metrics were used to analytically estimate the standard deviations of evaluation metrics using the following equation, where p is the

probability of condition being true and N is the number of trials. Binomial distribution is assumed.

$$SD = \sqrt{\frac{p * (1-p)}{N}}$$

Analytical estimate for standard deviation of gross conversion is 0.0202 and for standard deviation of net conversion it is 0.0156. With large samples binomial distribution resembles normal distribution, and there is not reason to expect either of the evaluation metrics would have a 'weird' shaped distribution. Both evaluation metrics are dependent on number of unique cookies that click "start free trial" button, making them coherent with the unit of diversion. Therefore analytic estimate is likely to be comparable to empirical variability.

Sizing

Analytical estimates of standard deviations were used to estimate the total number of page views needed to achieve sufficient statistical power. Alpha was chosen to be 0.05 and beta 0.2. A web-based Sample Size Calculator (http://www.evanmiller.org/ab-testing/sample-size.html) was used in calculations. The table below presents the results.

	Gross conversion	**Net conversion**
Beta	0.20	0.20
Alpha	0.05	0.05
Baseline conversion rate	0.2063	0.1093
Minimum detectable effect	0.01	0.0075
Sample size per branch	25835	27413
Sample size divided by click-through-probability	322937.5	342637.5
Total number of page views required	645875	685275

The total number of page views required to achieve sufficient power in both evaluation metrics is 685275. Based on the baseline metrics, the experiments needs to run for 18 days to gather such a number of page views. All the traffic is directed to the experiment. Bonferroni correction is not used due to the nature of the experiment: because both evaluation metrics need to be valid, the Bonferroni correction would be overly conservative. The experiment is not very risky for Udacity, because it only involves displaying a simple self-assessment question to some of the students. The experiment is not ethically problematic, and expected effects are limited to slight changes in conversion rates. Changes to user experience or site functionality (except for possible changes in student expectations) are minimal. Directing all the website traffic to the experiment should be safe for Udacity.

Experiment analysis

Sanity checks

As a first analysis step it is necessary to ensure the experiment and control conditions are comparable. This is done by comparing the numbers of cookies and clicks on "start free trial" button across the conditions. Given that 50 % of users should be directed to each condition, the 95 % confidence interval for invariant metrics was calculated assuming a binomial distribution. The table below shows the results.

Invariant metric	Lower bound	Upper bound	Observed	Passes
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks on "Start free trial"	0.4959	0.5041	0.5005	Yes

The observed values of both invariant metrics were within the 95 % confidence interval. As a result the sanity checks passed and the analysis can continue.

Result analysis

Effect size tests on the results were calculated using R (see the attached R script). The following table presents the results of effect size tests with 95 % confidence intervals. As hypothesized, the gross conversion was both statistically and practically significantly lower than in control condition, while the net conversion did not change significantly.

Evaluation metric	Lower bound	Upper bound	Statistical significance	Practical significance
Gross conversion	-0.0291	-0.0120	Yes	Yes
Net conversion	-0.0116	0.0019	No	No

Sign tests for the results were calculated using an online Sign and Binomial Test Calculator (http://graphpad.com/quickcalcs/binomial1.cfm). The results are shown in the table below.

	Number of days	Number of days when experiment > control	Baseline probability	Two-tailed p-value	Statistically significant
Gross conversion	23	4	0.50	0.0026	Yes
Net conversion	23	10	0.50	0.6776	No

According to sign tests the gross conversion was higher in experiment condition than in control condition on significantly fewer ($p < 0.05$) days than vice versa. In net conversion there was no significant difference.

Summary

The results of effect size and sign tests agree with each other. In the experiment condition the gross conversion was lower than in control condition. The net conversion decreased too, but not by a statistically significant amount. However, the confidence interval for net conversion includes the practically significant

decrease. Therefore the questionnaire might decrease net conversion by an amount that is large enough to affect the business. Bonferroni correction was not used in the test because all evaluation metrics need to be valid: the gross conversion should show a practically significant increase and at the same time the net conversion should not decrease significantly. Using the Bonferroni correction in this case would have been overly conservative.

Recommendation

The experiment results show that the self-assessment questionnaire decreased gross conversion by a practically significant amount. Net conversion decreased too, but not by a statistically significant amount. This means that students tend to make better-informed decisions on starting the free trial: those who decide to start the free trial continue it more often. However, at the same time there is a risk that the questionnaire might turn away some students that would have continued the course after the free trial period. Results of the experiment are therefore mixed. We can be confident that the implementation of self-assessment questionnaire makes a practically significant improvement in user experience of Udacity students, but we cannot rule out a possible negative effect on net conversion rate. My recommendation is to consider the decision to launch as a judgment call between user experience and Udacity business goals: if user experience is more important, the questionnaire should be launched, but if the financial factors are considered more important, the questionnaire should not be launched.

Follow-up experiment

After establishing that helping the students to set up their expectations in the beginning of the free trial with a simple questionnaire on their expected weekly commitment to learning decreases gross conversion without significantly affecting the net conversion rate, it would be interesting to find out whether the retention rate (number of user-ids that remain enrolled at least 14 days and make at least one payment, divided by number of user-ids to complete checkout) of those students that decide to try the free trial could be improved. People might be inaccurate in their evaluations of how much effort they have spend on learning during the week, and get frustrated if they believe they have been working more than they planned. In the follow up experiment the effect of displaying weekly effort to students on the website could be investigated. The experiment group would see a small display in the corner of the website showing the number of hours they have spent learning this week, and the control group would see a website without the display. The hypothesis is that tracking and displaying weekly effort to students would decrease the frustration, because the students could objectively evaluate if they have spent “too much” time trying to learn the content. As a result the retention rate of the experiment group should be higher than that of the control group.

Diversion to this experiment would happen after the students have provided their credit card information and completed the checkout to the free trial based on the user ids. The unit of diversion is thus the user-ids of user who enroll in the free trial. This choice on diversion allows the comparison of retention rates of

control and experiment groups, as the retention metric is dependent on number of checkouts. Invariant metric of the experiment would be number of user ids enrolling to the free trial, which makes it possible to check that equal number of users end up in the both conditions and diversion thus works as expected. Evaluation metric of the experiment would be retention rate. If significantly more students stay on the course after the 14-day trial in the experiment condition than in control condition, we can be confident that displaying the weekly effort students have put in learning helps to keep students engaged on the course.