A/B Testing Udacity's Free Trial Screener Project

Jocelyn (Yuan) Li

December 2016

1. EXPERIMENT OVERVIE

At the time of this experiment, Udacity courses currently have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

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 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. 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 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.

1.1 Metric Choice

- **Number of cookies:** That is, number of unique cookies to view the course overview page. (d_{min}=3000)
- **Number of user-ids:** That is, number of users who enroll in the free trial. (d_{min}=50)
- **Number of clicks:** That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (d_{min}=240)
- **Click-through-probability:** That is, 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:** That is, 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:** That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. (d_{min}=0.01)
- **Net conversion:** That is, 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. (d_{min}= 0.0075)

1.1.1 Invariant Metrics

There are two kind of metrics that measure how the experiment group is better than control. These are invariance metrics and evaluation metrics.

Invariant metrics is performing some kind of sanity check for us before running the experiment, like checking whether the distribution is the same. It performing some consistent checking across all of our experiments, which is why it shouldn't be changed. Because the new screen pops up after clicking on the 'start free trial' button, then the number of cookie, number of clicks and CTP should remain unchanged. Therefore, the following metrics was chosen for invariant metrics.

- Number of cookies
- Number of clicks
- Click-through-probability

1.1.2 Evaluation Metrics

Evaluation metrics, is usually the business metrics, like for example market share, number of users, or user experience metrics. Evaluation metrics are expected to change over the experiment. By comparing the different result between the control and experimental groups, we can then measure the effect of the new screener and test our hypothesis. Therefore, the following metrics was chosen for evaluation metrics. The following metrics are good evaluation metrics because they are directly dependent on the effect of the experiment.

- Gross conversion
- Retention
- Net conversion

1.2 Measuring Standard Deviation

We have collected the daily values for page views, cookies, enrollment rate, CTP, gross conversion, retention, and net conversion rate on Udacity's website. The number of clicks and enrollments follows a binomial distribution. For Bernoulli distribution with probability p and population N, the analytical standard deviation is computed as std = sqrt (p * (1-p) / N).

In the experiment, we predict that we will need approximately 5,000 cookies per day in each group.

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

- **Gross conversion**

The baseline probability for gross conversion is p = 0.20625, and the number of users who see the "start free trial" page (the denominator of the gross conversion) is N = 5000/40000*3200 = 400. Therefore the standard deviation is std = sqrt (p * (1-p) / N) = 0.0202.

The unit of analysis here is a person who click the "start free trial" page, and the unit of diversion is a cookie that does so. They are highly correlated, but not exactly the same.

- **Retention**

The baseline probability for retention is p = 0.53, and the number of users who enroll in the course is N = 5000/40000*660 = 82.5. Therefore the standard deviation is std = sqrt (p * (1-p) / N) = 0.0549.

The unit of analysis here is a person who enrolled the free trial, and the unit of diversion is the user-id that does so. This two almost always match up.

- Net conversion

The baseline probability for net conversion is p = 0.1093125, and the number of users who see the "start free trial" page is N = 5000/40000*3200 = 400. Therefore the standard deviation is std = sqrt (p * (1-p) / N) = 0.0156

The unit of analysis here is a person who click the "start free trial" page, and the unit of diversion is a cookie that does so. They are highly correlated, but not exactly the same.

1.3 Sizing

1.3.1 Number of Samples vs. Power

To know the exact number of pageviews required for our experiment, I will use an alpha value of 0.05 and beta value of 0.2 to calculate the pageviews. I used this calculator to determine how many page views we'll need to collect in our experiment.

- Gross conversion

Baseline Conversion: 20.625%

Minimum Detectable Effect: 1%

Alpha: 5%

Beta: 20%

1 - beta: 80%

Unique clicks/ unique pageviews=3,200/40,000=0.08

---> Sample Size=25,835

--->page views=25,835/0.08*2= **645,875**

- Retention

Baseline Conversion: 53%

Minimum Detectable Effect: 1%

Alpha: 5%

Beta: 20%

1 - beta: 80%

Enrollment/ unique pageviews=660/40,000=0.0165

---> Sample Size=39,155

--->page views=39,155/0.0165*2= **4,741,212**

- Net conversion

Baseline Conversion: 10.9313%

Minimum Detectable Effect: 0.75%

Alpha: 5%

Beta: 20%

1 - beta: 80%

Unique clicks/ unique pageviews=3,200/40,000=0.08

---> Sample Size=27,413

--->page views=27,413/0.08*2= **685,325**

The largest sample size is our limiting factor (retention rate), so we require a total of 4,741,212 pageviews to conduct the experiment

1.3.2 Duration vs. Exposure

If we divert 100% of traffic, given 40,000 page views per day, the experiment would take 4,741,212/40,000=119 around 119 days to finish. If we ignore retention and only use Gross Conversion and Net Conversion, then the experiment would take around 685,325/40,000 = 18 days.

A 119 day experiment with 100% diversion of traffic presents both a business risk (potential for: frustrated students, lower conversion and retention, and inefficient use of coaching resources) and an opportunity risk (performing other experiments). However, in general, this is not a risky experiment as the change would not be expected to cause a precipitous drop in enrollment. In terms of timing, an 18 day experiment is more reasonable, but % diversion may be scaled down depending on other experiments of interest to be performed concurrently.

2. EXPERIMENT ANALYSIS

2.1 Sanity Checks

For the sanity checks, I will check whether the invariant metrics are equivalent between the two groups. You can see the data from here.

- Number of cookies

Control Group Total Pageviews: 345,543

Experiment Group Total Pageviews: 344,660

Standard Deviation: sqrt(0.5*0.5/ (345,543+344,660))= 0.0006018

Margin of error: 1.96 * 0.0006018=0.0011796

Lower bound: 0.5 - 0.0011797 = 0.4988

Upper bound = 0.5 + 0.0011797 = 0.5012

Observed =345,543/ (345,543+344,660) =0.5006

The observed value is within the bounds, and therefore this invariant metric passed the sanity check.

- Number of clicks

Control Group Total Pageviews: 28,378

Experiment Group Total Pageviews: 28,325

Standard Deviation: sqrt(0.5*0.5/ (345,543+344,660))= 0.0021

Margin of error: 1.96 * 0.0021=0.004116

Lower bound: 0.5 - 0.004116= 0.4958840

Upper bound = 0.5 + 0.004116= 0.504116

Observed =28,378/ (28,378 +28,325) = 0.50047

The observed value is within the bounds, and therefore this invariant metric passed the sanity check

- Click-through-probability

For click through probability, we first compute the control value in control group and then compare the value in experiment group.

Control group CTP: 28,378/345,543= 0.082126

Standard Deviation: sqrt(0.082126*(1-0.082126)/ 345,543)= 0.0004671

Margin of error: 1.96 * 0.0004671=0.00092

Lower bound: 0.082126- 0.00092= 0.0812

Upper bound: 0.082126 + 0.00092= 0.0830

Experiment group CTP =28,325/344,660 = 0.082182

The experiment value is within the bounds, and therefore this invariant metric passed the sanity check.

All the invariant metrics passed the sanity checks.

2.2 Result Analysis

2.2.1 Effect Size Tests

For each evaluation metrics, I will calculate a confidence interval for the difference between the experiment and control groups, and check whether each metric is statistically and/or practically significance.

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

- 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.

X_cont= 3,785

N_cont=17,293 (sum of clicks for 23 days)

X_exp=3, 423

N_exp=17,260

p_pooled=(X_cont+ X_exp)/(N_cont+ N_exp)=(3,785 +3, 423)/(17,293 + 17,260)=0.2086

se_pooled = sqrt(p_pooled * (1-p_pooled) * (1./N_cnt + 1./N_exp))=sqrt(0.2086*(1-0.2086)*(1/17293+1/17260))=0.00437

margin of error=1.96* se_pooled=1.96*0.00437= 0.0086

d= X_exp / N_exp -X_cont/ N_cont=3, 423/17,260-3,785/17,293=-0.02055

d_min = 0.01

CI (-0.029, -0.012)

The confidence interval does not include 0, so the gross conversion is statistically significant; also, the confidence interval does not include the practical significance boundary, so the gross conversion is also practically significant.

- **Net conversion**

Gross 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.

X_cont= 2,033

N_cont=17,293 (sum of clicks for 23 days)

X_exp=1,945

N_exp=17,260

p_pooled=(X_cont+ X_exp)/(N_cont+ N_exp)=(2,033 +1,945)/(17,293 + 17,260)= 0.1151

se_pooled = sqrt(p_pooled * (1-p_pooled) * (1./N_cnt + 1./N_exp))=sqrt(0.1151*(1-0.1151)*(1/17293+1/17260))=0.00343

margin of error=1.96* se_pooled=1.96*0.00343= 0.00672

d= X_exp / N_exp -X_cont/ N_cont=1,945/17,260-2,033/17,293=-0.0048

d_min = 0.0075

CI (-0.0116, 0.0019)

The confidence interval include 0, so the net conversion is not statistically significant; also, the confidence interval includes the practical significance boundary, so the gross conversion is also not practically significant.

2.2.2 Sign Tests

I use this online calculator to perform sign test.

- **Gross conversion**

Control				Experiment			
Clicks	Enrollments	gross conversion		Clicks	Enrollments	gross conversion	
687	134	0.195050946		686	105	0.153061224	0
779	147	0.188703466		785	116	0.147770701	0
909	167	0.183718372		884	145	0.164027149	0
836	156	0.186602871		827	138	0.166868198	0
837	163	0.19474313		832	140	0.168269231	0
823	138	0.167679222		788	129	0.163705584	0
748	146	0.195187166		780	127	0.162820513	0
632	110	0.174050633		652	94	0.144171779	0
691	131	0.189580318		697	120	0.172166428	0
861	165	0.191637631		860	153	0.177906977	0
867	196	0.226066897		864	143	0.165509259	0
838	162	0.193317422		801	128	0.15980025	0
665	127	0.190977444		642	122	0.190031153	0
673	220	0.326894502		697	194	0.278335725	0
691	176	0.254703329		669	127	0.189835575	0
708	161	0.22740113		693	153	0.220779221	0
759	233	0.306982872		771	213	0.276264591	0
736	154	0.20923913		736	162	0.220108696	1
739	196	0.265223275		727	201	0.27647868	1
734	167	0.227520436		728	207	0.284340659	1
706	174	0.246458924		722	182	0.252077562	1
681	156	0.22907489		695	142	0.204316547	0
693	206	0.297258297		724	182	0.251381215	0

For gross conversion, the number of days we see an improvement in experiment group is 4, out of total 23 days of experiment. The two-trail p value is 0.0026, which is smaller than alpha 0.05. Therefore, the change is statistical significant.

- Net conversion

Control			Experiment			
Clicks	payments	net conversion	Clicks	payments	net conversion	
687	70	0.101892285	686	34	0.049562682	0
779	70	0.089858793	785	91	0.115923567	1
909	95	0.104510451	884	79	0.089366516	0
836	105	0.125598086	827	92	0.111245466	0
837	64	0.07646356	832	94	0.112980769	1
823	82	0.09963548	788	61	0.077411168	0
748	76	0.101604278	780	44	0.056410256	0
632	70	0.110759494	652	62	0.095092025	0
691	60	0.08683068	697	77	0.110473458	1
861	97	0.112659698	860	98	0.113953488	1
867	105	0.121107266	864	71	0.082175926	0
838	92	0.109785203	801	70	0.087390762	0
665	56	0.084210526	642	68	0.105919003	1
673	122	0.18127786	697	94	0.134863702	0
691	128	0.185238784	669	81	0.121076233	0
708	104	0.146892655	693	101	0.145743146	0
759	124	0.163372859	771	119	0.154345006	0
736	91	0.123641304	736	120	0.163043478	1
739	86	0.116373478	727	96	0.132049519	1
734	75	0.102179837	728	67	0.092032967	0
706	101	0.14305949	722	123	0.170360111	1
681	93	0.136563877	695	100	0.143884892	1
693	67	0.096681097	724	103	0.142265193	1

For net conversion, the number of days we see an improvement in experiment group is 10, out of total 23 days of experiment. The two-trail p value is 0.6776, which is larger than alpha 0.05. Therefore, the change is not statistical significant.

2.3 Summary

I decide not to use Bonferroni correction, because the metrics in the test has high correlation and the Bonferroni correction will be too conservative to it.

Both the effective size hypothesis tests and sign tests state that the change will practically significantly reduce the gross conversion, but not affect the net conversion rate in a practically significant ways.

3. RECOMMENDATION

Based on the analysis above, I recommend not to adopt the changes of adding "5 or more hour" recommendation to "start free trial" date. The reason is that the A/B test shows that this will not practically significantly increase the net conversion rate. In other words, it does not increase the number of paid users, which fails the original goal of launching this feature.

4. FOLLOW-UP EXPERIMENT

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.