A/B Testing project

Alain Roghi - Airbus

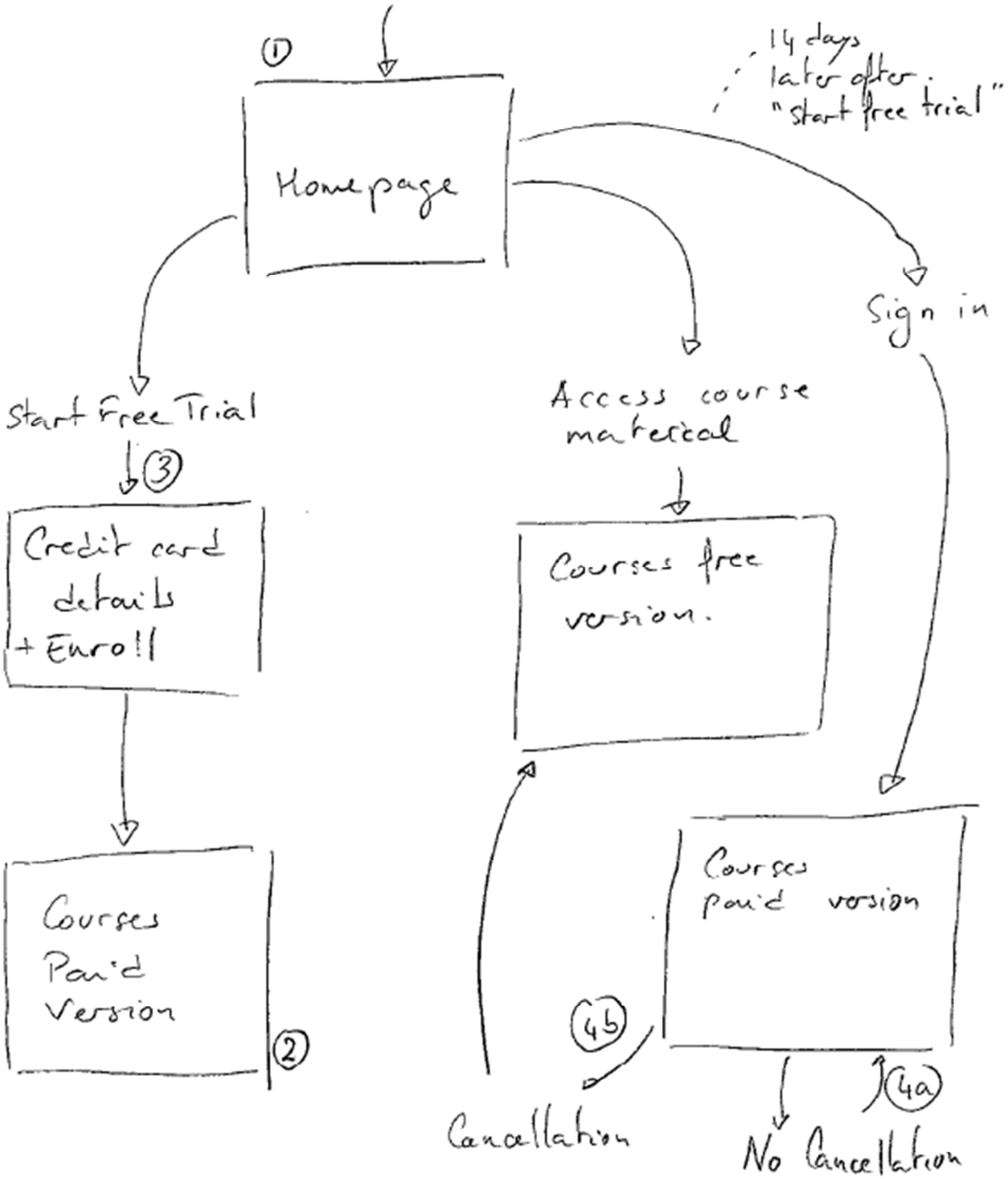
Business understanding

The aim of this paragraph is to provide a business understanding of the Udacity proposed changes and the expected impact Udacity expect.

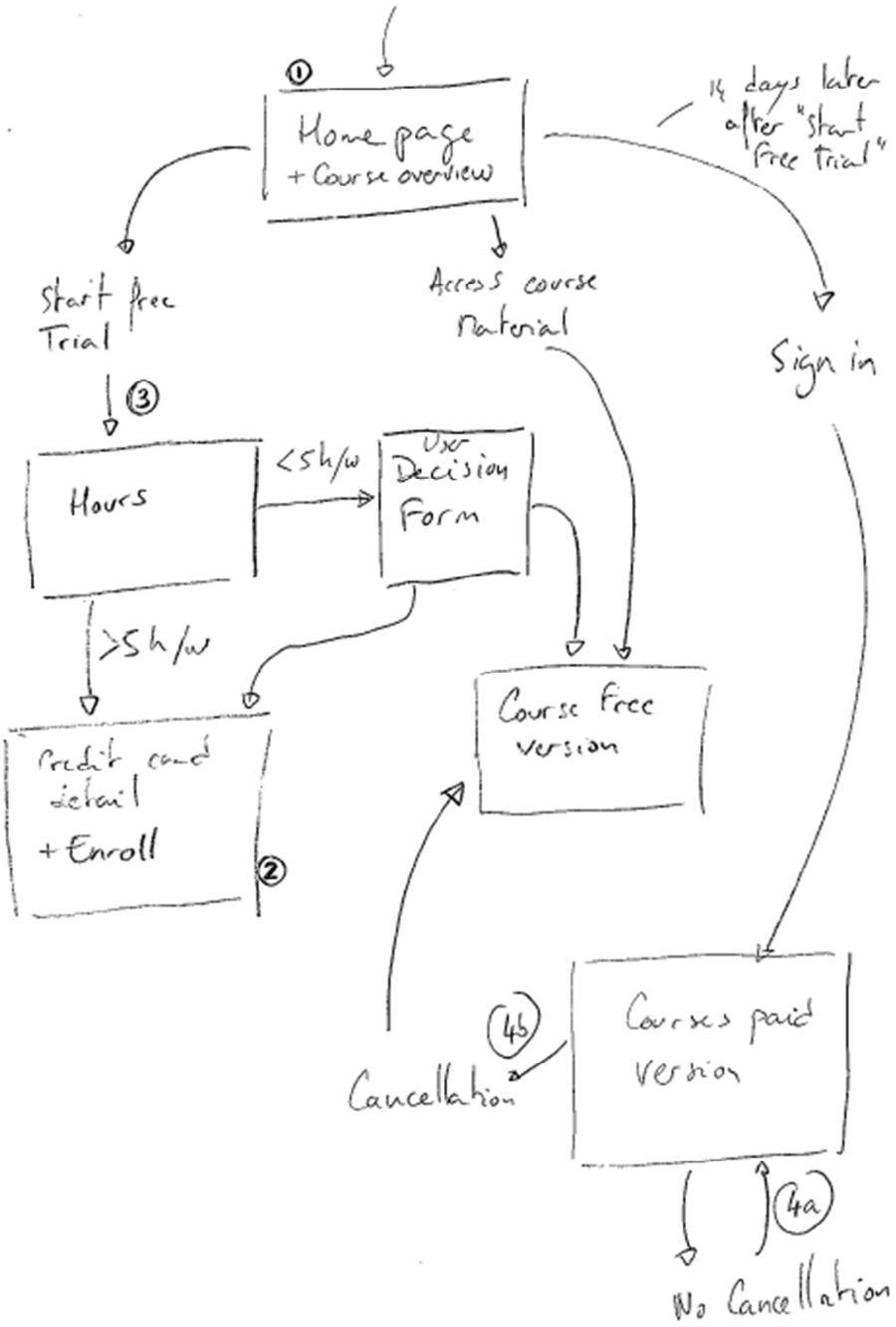
For this, I use a user experience flow diagram of the current Udacity website, and the proposed modified one.

The different figures on the diagram show points of measures to capture raw metrics.

Current user experience flow diagram:



New proposed user experience flow diagram:



From Udacity expectation and changes done, we do not expect changes in the number of people visiting the Udacity homepage and courses list. There should not be any change in term of people clicking the ‘Start free trial’ button.

By challenging the user enrolment, Udacity expect to have less user enrolled for the 14 days free trial, but more engaged peoples. At the same time, Udacity expect less cancellation at the

end of the 14 days free trial period. In other words, Udacity expect to change the 4a/4b distribution by Increasing 4a (in %) and reduction 4b (in %).

Experiment Design

Metric Choice

The following table maps the proposed metrics on the user experience flow diagram. I also try to get an intuition about the change direction that should occur on the metric if the experiment is successful:

Metrics / definition	Place of data collection or computation formula	Possible change / explanation
Number of cookies number of unique cookies to view the course overview page	1	No change The Udacity planned change shall have no effect on the number of internet user visiting the homepage and course overview page. This metric can be used as an invariant metric. If we notice significant changes, it may be due to a failure in the experiment.
Number of userids number of users who enroll in the free trial.	2	Decrease By challenging user decision, we can expect more users going to free section, and only the more engaged ones starting the 14 days trial period. Nevertheless,
Number of clicks number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger)	3	No change The Udacity planned change shall have no effect on the number of internet user clicking on the 'Start free trial' button. This metric can be used as an invariant metric. If we notice significant changes, it may be due to a failure in the experiment.
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	3 / 1	No change As the underlining collected metrics shall not change, we are not expecting any change on that metric. This metric can be used as an invariant metric. If we notice significant changes, it may be due to a failure in the experiment.
• Gross conversion number of userids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button.	2 / 3	Decrease As the number of created user ids (2) could decrease, maintaining the number of clicks, we can expect a decrease of the gross conversion.
• Retention	4a / 2	Increase

Metrics / definition	Place of data collection or computation formula	Possible change / explanation
number of userids to remain enrolled past the 14day boundary (and thus make at least one payment) divided by number of userids to complete checkout		As mentioned earlier, for the same number of user id, we expect an increase of people none cancelling the paying service. Therefore, we expect the retention metrics to increase.
• Net conversion number of userids to remain enrolled past the 14day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button	4a / 3	Decrease We expect retention to increase, but it may be at the end a lower number of users in absolute value. As the number of cookies clicking on the 'Start free trial' button shall be stable, we expect the Net conversion to decrease. **Udacity does not expect decrease of this metric and even would like this metric to increase.**

As the 'number of user id' metric and the 'Gross conversion' are highly correlated, I will only select one of these 2 metrics as evaluation metrics. As the 'number of user id' is a measure and is meaningless without a comparison baseline and the 'Gross conversion' is a metric including this comparison baseline, I would recommend using the 'Gross conversion'.

As Udacity wants to improve user experience, it shall be reflected by within the 'retention' metric. This shall be one of our evaluation metrics.

As Udacity do not expects to reduce the number of people to continue past the free trialperiod and an improved user experience, this shall be reflected in the 'Net conversion' metric. This shall be one of our evaluation metrics.

I propose to use the following metrics as invariant metrics for sanity check:

- Number of cookies
- Number of clicks
- Click through rate

I propose to use the following metrics as evaluation metrics:

- Gross conversion
- Retention
- Net conversion

In order to accept the modification we expect:

- The Gross conversion to decrease
- AND the retention to increase
- AND the net conversion to be stable or increase

The gross conversion metric addresses the “his 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” part of the initial hypothesis.
The retention and net conversion metrics address the “without significantly reducing the number of students to continue past the free trial and eventually complete the course” part of the initial hypothesis.

Measuring Standard Deviation

The provided figures are:

Unique cookies to view page per day:	40000
Unique cookies to click "Start free trial" per day:	3200
Enrollments per day:	660
Click-through-probability on "Start free trial":	0.08
Probability of enrolling, given click:	0.20625
Probability of payment, given enroll:	0.53
Probability of payment, given click	0.1093125

If we assume only 5000 unique cookies to view per per day, we will get the following figures:

Unique cookies to view page per day:	5000	
Unique cookies to click "Start free trial" per day:	400	= 3200 * 5000 / 40000
Enrollments per day:	82.5	= 660 * 5000 / 40000
Click-through-probability on "Start free trial":	0.08	
Probability of enrolling, given click:	0.20625	
Probability of payment, given enroll:	0.53	
Probability of payment, given click	0.1093125	

We get the following analytic standard deviation for the 3 evaluation metrics:

Metric	Standard deviation	Formula
Gross conversion	0.0202	=sqrt(0.20625*(1-0.20625)/400)
Retention	0.0549	=sqrt(0.53*(1-0.53)/82.5)
Net conversion	0.0156	=sqrt(0.1093125*(1-0.1093125)/400)

Our test has the cookie as unit of diversion.

For the Retention metrics, denominator is user-id. So, I expect to have a much higher empirical variability.
For the Gross conversion and Net conversion, the denominator matches the unit of diversion of the test. In that case, I expect to have a higher empirical variability close to the analytical one.

Sizing

Number of Samples vs. Power

As we are looking to 3 different evaluation metrics with a confidence level of 95% ($\alpha = 0.05$), the probability to have a false positive is 1-0.95*0.95*0.95 = 0.14. I consider it is still acceptable and would not recommend using the Bonferroni correction.

To compute the number of needed pageviews, Ewans Miller online calculator available at: http://www.evanmiller.org/ab-testing/sample-size.html.

Metric	Baseline conversion rate	Minimum Detectable Effect	Population in each group	Number of page views
Gross conversion	20.265%	1%	25502 clicks	637550 25502 * 2 / 3200 * 40000
Retention	53%	1%	39087 User ids	4737818 39087 * 2 / 660 * 40000
Net conversion	10.93125%	0.75%	27413 clicks	685325 27413 * 2 / 3200 * 40000

$\alpha = 0.05$
$\beta = 0.2$

To be able to conduct our analysis, we will need to have more than 4.737.818 pages view in our test.

Risk assessment

We fundamentally do not change the registration process. The new process does not change in any way the information collected for each participant nor the price he will pay. We can consider that this experiment does not generate any risk greater than minimal risk for the participant.

Duration vs. Exposure
We have currently 40000 pages view per day. It means if we divert 100% of the traffic to the test, it will last around 119 days, so around 16 weeks. This is a very long duration with a 100% traffic diversion. I think this proposal cannot be made to Udacity deciders.

Therefore, I propose not to use the Retention as an evaluation metric anymore and to rely only on the Gross and Net conversion metrics. As I only have 2 evaluation metrics, use of the Bonferroni correction even less needed.

In mean that I will “only” need 685325 pages view for that study. If we divert 100% of the traffic to the test, it will take 18 days. In addition, we need to wait another 14 days in order end the full experiment process (up to cancellation or payment). It means a total of 32 days.

As the experiment is not risky for the participant, I suggest to divert 100% of the traffic.

Experiment Analysis

Sanity Checks

As the invariant metrics are not related to enrolment nor payment, we can use the full set of data for the provided 37 days.

Control set

Total pages view	Total click	Click trough probability
345543	28378	0.082126

Experiment set

Total pages view	Total click	Click trough probability
344660	28325	0.082182441

Total numbers

Total pages view	Total click
690203	56703

For the pages view and total click metrics, we have the following figures:

Metric	Standard error	Lower CI bound	Upper CI bound	Observed value
Pages view	0.000602	0.4988	0.5012	0.5006
Click	0.0021	0.4959	0.5041	0.5005

For these 2 metrics, the observed value is within the lower and upper CI bound.

For click through probability, we have:

- pooled probiblity = 0.0822
- d hat = 0.082182441 - 0.082126 = 0.0001 (rounded 4 decimal places).

Standard error is 0.000661 (= sqrt(0.0822*(1-0.0822)/(1/345543+1/344660)))
Confidence interval is: +/- 1.96*0.000661 = +/- 0.0013
The d hat value is within the confidence interval.

All the 3 invariant metrics passed the sanity check.

Result Analysis

Effect Size Tests

For the two evaluation metrics, with figures computed with values up to Nov 2, we get the following results:

Metric	Standard error	Lower bound	Upper bound	D Hat	Pooled probability
Gross conversion	0.004372	-0.0291	-0.0120	-0.020554875	0.208607067
Net conversion	0.003434	-0.0116	0.0019	-0.004873723	0.115127485

The gross conversion metrics is statistically and practically significant (d min = 0.01).

The net conversion metrics is not statistically nor practically significant (d min = 0.0075).

Sign Tests

The gross conversion metrics was greater 4 time in the experiment vs the control group (over 23 days).

The net conversion metrics was greater 10 time in the experiment vs the control group (over 23 days).

Using the online calculator http://graphpad.com/quickcalcs/binomial1.cfm with these values, I get a 2 tails p probability of:

Metric	two-tail P value
Gross conversion	0.0026
Net conversion	0.6776

So, decrease of the gross conversion seems not to be due by chance, but change of the net conversion is likely due to change.

Summary

Our decision to go or not to go for the modification will be based on having a decrease of the gross conversion AND having the net conversion stable.

It means that we will reject the modification if at least one of the metrics does not fulfil our test.

If we state our null hypothesis is:

- H0 : gross conversion is stable or increase and the net conversion decrease

And our alternative hypothesis is:

- Ha : gross conversion decrease and net conversion is stable or increase

In our case, our risk is to accept the modification (meaning to reject the null hypothesis) while the null hypothesis is true. It will have financial impact on Udacity. This is a type I error.

We set the significance level to 5% for both the gross and net conversion metrics. In other words, the probability of observing a significant metric change by chance is 5%. Observing both metrics changes by chance at the same time is far less than 5%.

In other words, the probability of detecting by chance the null hypothesis is very low. And therefore the risk of rejecting the null hypothesis when the null hypothesis is true is already very low (Type 1 error). As the Bonferroni correction aims to reduce the risk of Type 1 error (by reducing the significance level for each metric), it would not be convenient to apply in our case.

As a conclusion, I recommend not to use the Bonferroni correction.

Source: https://generallythinking.com/what-the-hell-is-bonferroni-correction/https://en.wikipedia.org/wiki/Type_I_and_type_II_errors#Statistical_test_theory
https://en.wikipedia.org/wiki/Probability_of_error
http://ethen8181.github.io/Business-Analytics/ab_tests/frequentist_ab_test.html
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4333023/

Remark: P(A B) = P(A)xP(B) for independent events, P(A B) = P(A)xP(B|A) for dependent events.

But I can try to analyse metrics using the Bonferroni correction. It means I want now an confidence interval at 97.5% (2 metrics). Using this confidence interval (I used a z score of 3.03), and the new computed figures, I consider the gross conversion decrease still statistically significant but pragmatically less significant. I do not change the conclusion for the net conversion metrics.

I do not see any contradiction between the effect size hypothesis tests and the sign tests.

Recommendation

I saw our experiment was properly designed using sanity checks.

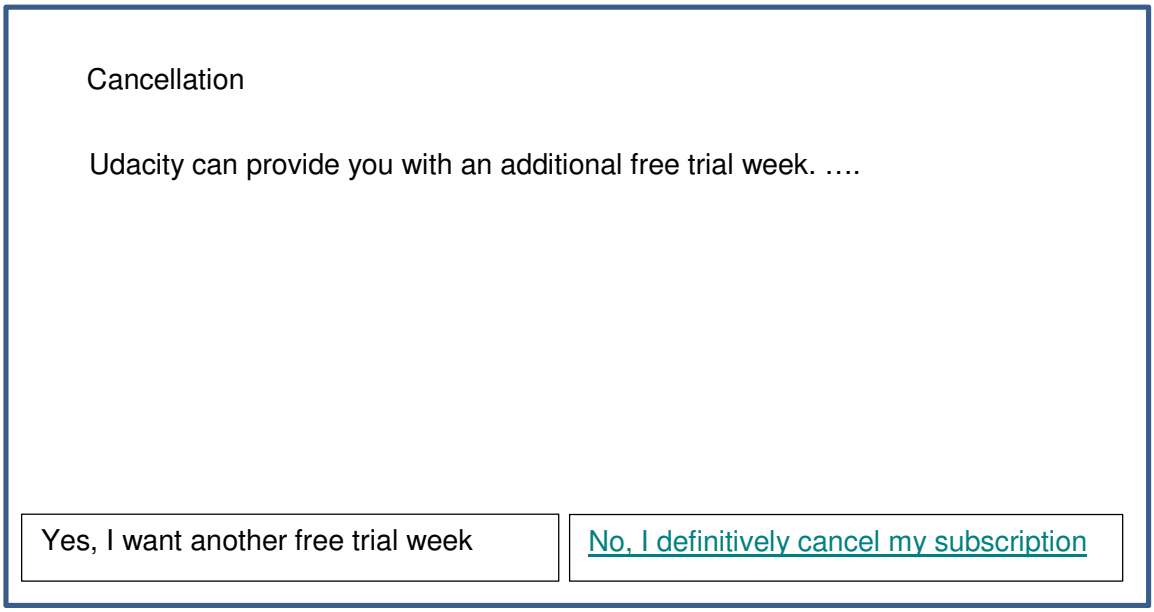
I saw a significant decrease of the gross conversion. This is the expected result for this metrics.

For the net conversion, we were expecting a stable or even an increasing value. We experience a decrease. This decrease is not statistically significant. Nevertheless, it does not go in the direction we wanted. Also, the lower bound of our confidence interval exceed the practical negative limit (-0.0075) and the $\hat{d}$ value is close from this practical negative limit (-0.0075). In other words, we start experiencing a practically significant decrease of payment and it can exceed the practical threshold set by Udacity.

With the current information we retrieve from this experiment, my recommendation would be to stop the experiment and not to go for this modification.

Follow-Up Experiment

The new proposed experiment aims to challenge the cancellation decision of the user and to give him the change to have one additional free trial week. The form could look like this:



At the end of the additional free trial week, student will be automatically enrolled unless if he decides to cancel. After the additional free trial week, the student does not get the option of having another free trial week.

The null hypothesis will be that Udacity won't get any change in student cancellation rate (or even an increased rate). The alternate hypothesis is that student cancellation rate will significantly decrease at the end of the 14 days but also (=and) after the full trial period.

The proposed metrics are:

Number of user-id: That is, number of users who enroll in the free trial. This metric will act as an invariant metric to perform sanity check on our experiment. We can also set a dmin value of 50.

Cancellation form click probability: That is, number of user-id to click the "Cancellation" button after the 14 days trial period divided by number of user-id enrolled in the trial period. This metric shall also act as invariant metric to perform sanity check on our experiment. We can also set a dmin value of 0.01.

Cancellation rate after the full trial period: That is:

- For the control group: the number of unique user-id to cancel after the 14 days free trial period divided by the number of user-id enrolled in the trial period.
- For the experiment group: the number of unique user-id to cancel either afther the first 14 days trial period or after the additional weel of trial period, divided by the number of user-id enrolled in the trial period.

This metric will act as evaluation metric and we can set a dmin of 0.0075.

From a business perspective, I expect to have a significant decrease of student cancelling subscription after the end of their free trial period.
In other words, I expect to retain some student in the free trial after 14 days, and to keep them enrolled after the 3 weeks trial period.

As my experiment starts only after student enrolment, I propose to have the user-id as unit of diversion. As the unit of analysis is also the user-id for the various metrics, I will reduce the metrics variability.