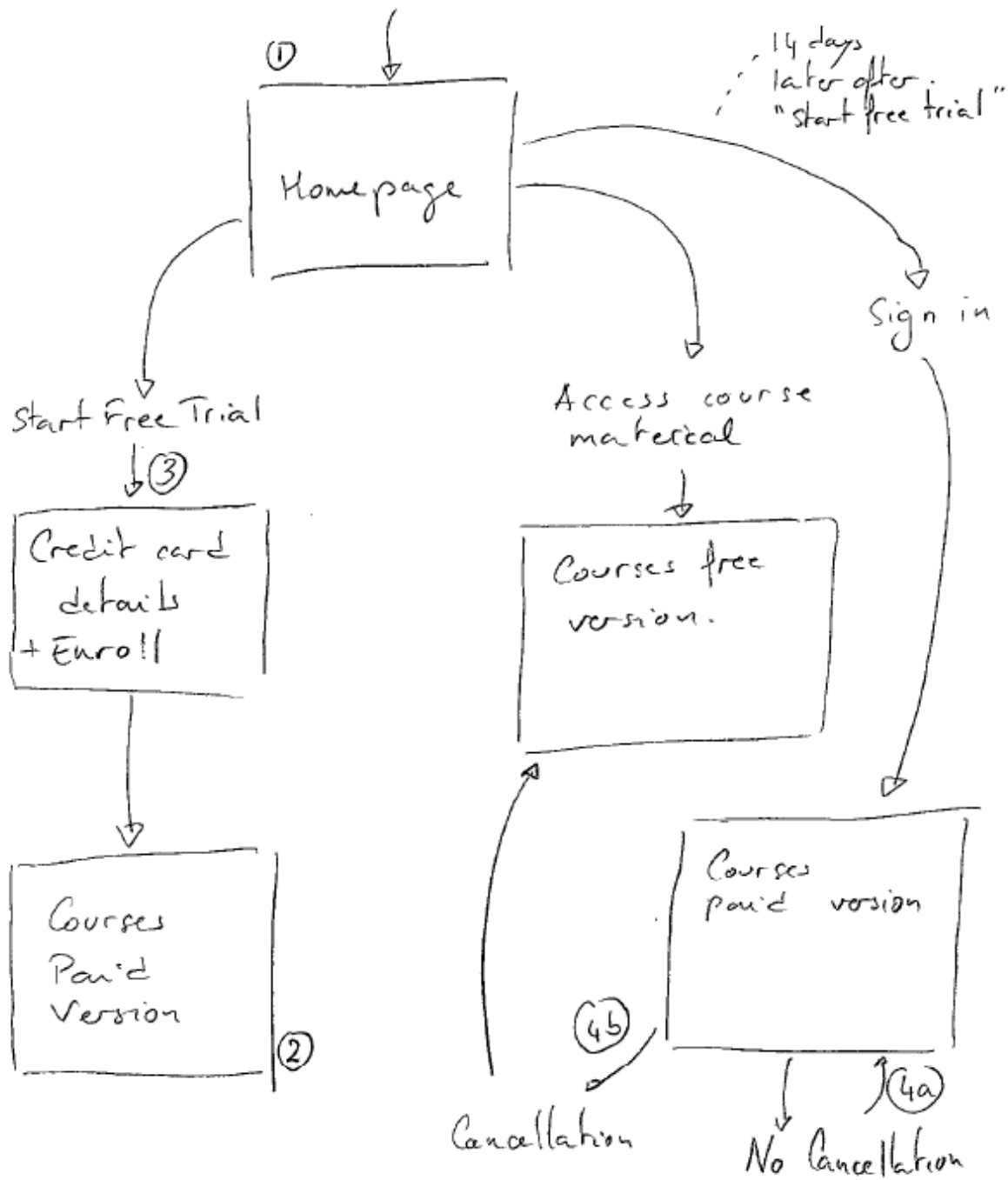# A/B Testing project

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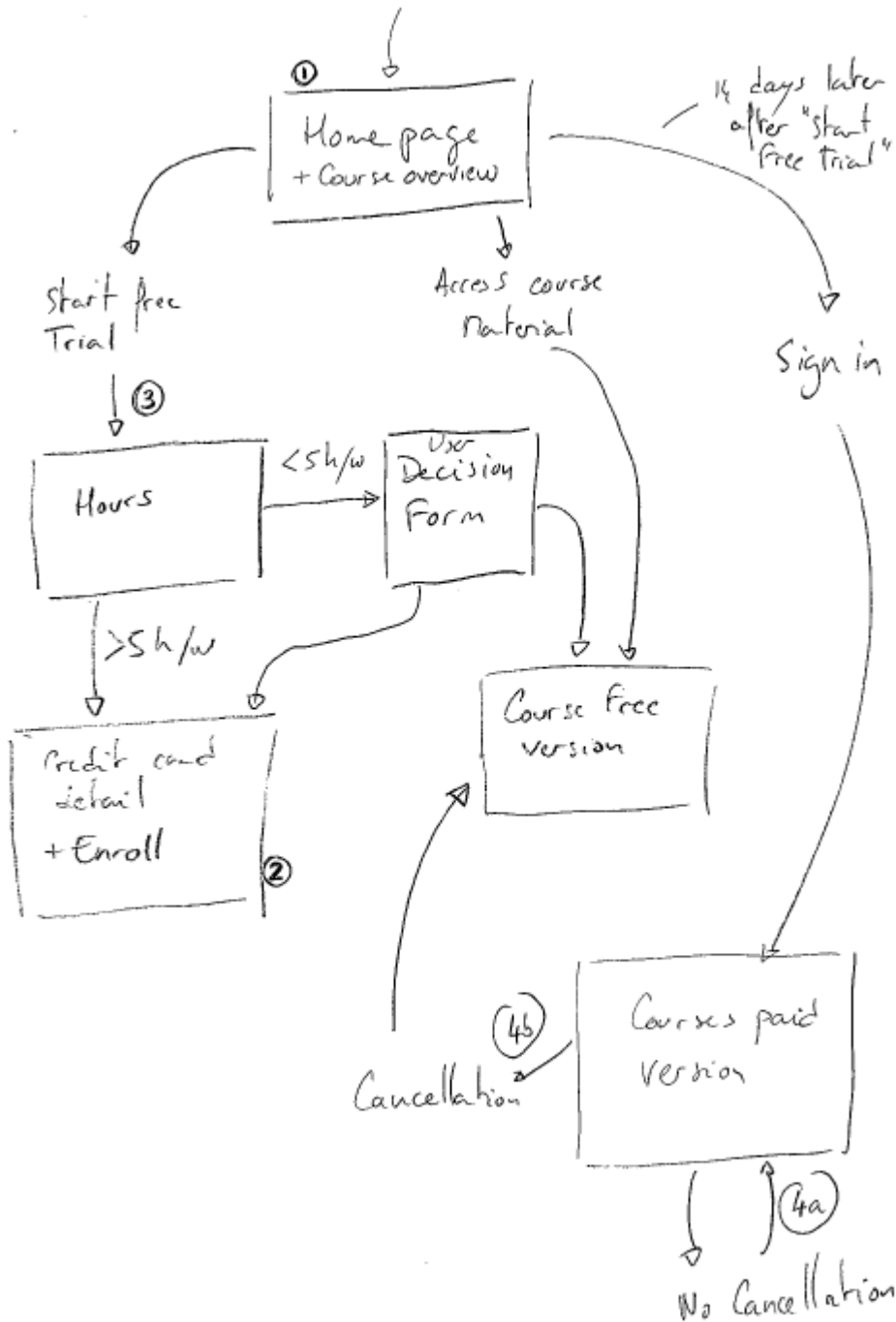
For this, I use a user experience flow diagram of the current Udacity website, and the proposed modified one.

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

Current user experience flow diagram:

① Homepage

"14 days later after 'start free trial'"

Start Free Trial

③ Credit card details + Enroll

Courses Paid Version ②

Access course material

Courses free version.

Sign in

Courses paid version

④b Cancellation

④a No Cancellation

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, **Udacity does not expect a significant modification of this metric.** This means Udacity expects a behavioural change from its users.
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. **Udacity does not expect a significant modification**

Metrics / definition	Place of data collection or computation formula	Possible change / explanation
		of this metric.
• Retention 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	4a / 2	Increase 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. **Udacity even expect the 'number of userid' metric to be stable or even to be bigger. It means that even if we get an increase, it could be less than initially thought.**
• 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 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 entering the free trial period 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

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% (α = 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

α = 0.05
β = 0.2

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

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.

The modification appears not to be risky and I propose to divert 50% of my traffic to the experiment. The experiment duration will last 35 days. As the latest entered user in the experiment will still have 14 days before deciding if they stay or not within the experiment, the total duration of my experiment will be 49 days. I think this is an acceptable duration to be proposed to Udacity deciders.

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 change, but change of the net conversion is likely due to change.

Summary

I did not used the Bonferroni correction as I considered that the probability to have a false positive was acceptable. I then reduced the number of evaluation metrics from 3 to 2. This empathise the fact I do not need the Bonferroni correction.

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

Make a recommendation and briefly describe your reasoning.

I saw our experiment was properly designed using sanity checks.

I saw a significant decrease of the gross conversion. Even if this decrease could be expected, this was not the Udacity desired effect. So behavioural change expected by Udacity has not occurred.

I saw a non-significant decrease of the net conversion. It means that the decrease of Udactity revenues may be due by chance. To be really rigorous, I would suggest Udacity to continue the test in order to reach the requested number of page views to have significant analysis of the experiment. Nevertheless, as the expected behavioural change has not occurred, there is a high risk to get at the end a significant decrease of the net conversion metrics.

My recommendation would be to stop the experiment and not to go for this modification.

Follow-Up Experiment

The initial objective was to decrease frustration while maintaining the level of enrolment and payment. I think our experiment as designed was not reaching this target. I suggest we could just redo the same experiment without any change but indicating a minimum of 4 hours per weeks instead of 5. This mean our experiment could be started rapidly. We could see the impact of the 4 hours indication versus the current behaviour. But we could also test the impact of reducing the indicated number of ours from 5 to 4 hours. It could help us to determine if hour is the right leveller we can trigger to decrease frustration.