

A/B TESTING PROJECT

MARIO BONILLA

May 4, 2016

OVERVIEW

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.

EXPERIMENT DESIGN

1. Metric Choice

The launch criteria should measure the two main goals of the experiment:

- Decrease the number of students who sign up for the free trial but don't continue past the free trial,
- Without decreasing the revenue.

In terms of metrics, to launch the experiment, we must observe a practically and statistically significant decrease in gross conversion as well as a practically and statistically significant not decrease for net conversion and retention.

In order to have a good understanding of the experiments results and/or perform comparisons to test the significance (statistical and practical), we have to choose both Invariant and Evaluation metrics:

- **Invariant metrics** are those that do not suffer changes (hence the name) between the control and experiments groups; in other words, we expect them to follow the same distribution in both groups. In this particular case, we have chosen the **number of cookies**, the **number of clicks**, and the **click-through-probability** as invariant metrics.

Invariant Metrics	
Number of Cookies number of unique cookies to view the course overview page	<ul style="list-style-type: none">- It is the unit of Diversion; normally, evenly distributed in groups (control and experiment).- independent from the experiment (good invariant, bad evaluation metric): clicks happen before the user interacts with the experiment.
Number of Clicks number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger)	<ul style="list-style-type: none">- independent from the experiment dependent from the experiment (good invariant, bad evaluation metric): clicks happen before the user interacts with the experiment.- At this point all the users have the same experience (same Course Overview Page), therefore, same distribution among groups (control, experiment) is likely to happen.
Click thru Probability number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.	<ul style="list-style-type: none">- independent from the experiment: clicks happen before the user interacts with the experiment.- At this point all the users have the same experience (same Course Overview Page), therefore, same distribution among groups (control, experiment) is likely to happen.

- **Evaluation metrics** could vary among the control and experiment groups; in other words, we foresee that the distributions could be different between groups as a consequence of the experiment. In this case we chose the **gross** and **net conversion** as well as the **retention** as evaluation metrics. We would like to see more student retention in the experiment group as well as more net conversion (our changes work for the better). It would be also positive to see a decrease in gross conversions (less costs), given an increase in net ones (more revenue), of course.

Evaluation Metrics	
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.	<ul style="list-style-type: none"> - dependent from the experiment (good evaluation, bad invariant metric): the metric could throw different results from the different groups and therefore we could compare them. The control group could enroll/pay without knowing the minimum time requirements, but the experiment group are shown the time questions and they have to make a decision (enroll-pay vs. continue free trial-not pay). Based on this, we could expect that the gross conversion is higher in the control group vs. the experiment group. We will evaluate that. - It must be evaluated if the minimum threshold value of 0.01 is or not reached.
Retention 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	<ul style="list-style-type: none"> - dependent from the experiment (good evaluation, bad invariant metric): the metric could throw different results and therefore we could compare them. The control group could enroll/pay without knowing the minimum time requirements, but the experiment group are shown the time questions and they have to make a decision. In this case, we could expect the retention to be higher in the experiment group vs. the control group. - It must be evaluated if the minimum threshold value of 0.01 is or not reached.
Net Conversion 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.	<ul style="list-style-type: none"> -The result from the numbers above (gross and retention) - dependent from the experiment (good evaluation, bad invariant metric): the metric could throw different results and therefore we could compare them. The control group could enroll/pay without knowing the minimum time requirements, but the experiment group are shown the time questions and they have to make a decision. - It must be evaluated if the minimum threshold value of 0.0075 is or not reached.

- **Not Used Metrics**
 In this case, we did not use the **number of user-ids** metric because we find it not to be neither a good invariant nor a good evaluation metric. It is not a good invariant one because users enrolled in the free trial is dependent on the experiment and it is not also a good evaluation metric because the number of visitors could vary between the experiment vs. control group.

2. Measuring Standard Deviation

We have measured the Standard Deviation first for 3200 clicks and 40000 page views (that number is needed for the next calculation). After that we calculated it for 5000 clicks and the same number or page views, 40000.

	Standard Deviation (for 5000 clicks & 40000 page views)
Gross Conversion	0.0202
Retention	0.549
Net Conversion	0.0156

The unit of diversion is the number of cookies; this figure is also used as denominator when calculating both the gross and net conversion. Because of that, the gross and net conversion could be used to make comparisons between the empirical variability and the analytical variability. The analytical variability of the standard deviation approximates the empirically calculated standard deviation when the unit of diversion is equal to the unit of analysis

However, the denominator for the retention calculation is the number of users who enroll in the course, therefore different from the unit of diversion. Therefore, it could be the case that the empirical variability is different from the analytical variability and then we will have to calculate both values.

3. Sizing

After some consideration we decided not to use the Bonferroni method for analysis.

In order to power appropriately our experiment, we have to give an optimal number of page views; in our experiment, the total number of page views is **685325**.

	Gross Conversion	Retention	Net Conversion
Baseline Conversion:	0.20625	0.53	0.109313
d min:	0.01	0.01	0.0075
α	0.05	0.05	0.05
β	0.20	0.20	0.20
Sample Size	25835	39155	27413
Number of Groups	2	2	2
Total Sample Size	51670	78230	54826
Ratio Clicks/Page View	0.08	0.0165	0.08
<u>Page Views</u>	645875	4741212	<u>685325</u>

Sample Size figure is the number of enrollments by group
Number of Groups: only two, control and experiment
Ratio Clicks/Page View: click thru probability

That number of viewed pages could be reached in **18 days** given the rate of 40000 view pages per day, with a **100% diversion**. We find that time frame financially cost effective, and it will not require it to be performed during a long period of time (a little bit less than three weeks), hence the risk will not be high.

Should other considerations yet unknown come into play, we could lower the rate of diversion to **75%** (in this case we would need **23 days** to reached the desired number of page views), or we could lower the diversion even further to **50%** (in this case **35 days** will be needed).




It is also true that we could have used the retention page views figure (almost 5 million viewed pages) as a baseline for our experiment but that would have expanded the experiment to about 120 days (about 3 months) if 100 % diversion used. That could have posed some financial drawbacks as well as some possible customer frustration because of the length of the experiment.

From the metrics at our disposal, we chose the net conversion because, having ruled out the retention for being too high, in our opinion, the net conversion is more difficult to achieve than the gross conversion under a business perspective.

EXPERIMENT ANALYSIS

4. Sanity Checks



We will test the invariant metrics to check if there is equal diversion in both the experiment and the control groups. We calculate the upper and lower interval bounds (at 95 % confidence) and check if the observed values lie within those interval limits. If so, the sanity check is passed.

	Number of Cookies	Number of clicks on "start free trial"	Click-through-probability
Expected Value	0.5	0.5	0.0821
Observed Value	0.5006	0.5005	0.0822
Confidence Interval (@ 95%)	[0.4988 – 0.5012]	[0.4959 – 0.5042]	[0.0812 – 0.083]
Pass Sanity Check			

All three invariant metric pass the Sanity Check, at 95 % confidence interval.



5. Result Analysis

For each of the evaluation metrics, we build a 95% confidence interval around the difference between the experiment and control groups and check whether each metric is statistically and practically significant.

	Gross Conversion	Net Conversion
d_{\min} (minimum)	0.01	0.0075
Observed Difference	-0.0205	-0.0048
Confidence Interval (@ 95%)	$[(-0.0291) - (-0.0116)]$	$[(-0.012) - (-0.0019)]$
Pass both Statistical and Practical Significance Test		

While the Gross conversion passes both the statistical and practical significance tests @ 95% confidence interval, the Net conversion fails both of them.

Also, for each of the evaluation metrics, we perform a **sign test** using the day-by-day data, and we report the p-value of the sign test and whether the result is statistically significant.

	Gross Conversion	Net Conversion
P value	0.0026	0.6776
Pass Statistical Significance Test $\alpha = 0.05$		

As shown in the table above, the Gross conversion metric passes the sign test (the metric is statistically and practically significant) and the Net conversion does not (is statistically and practically not significant), with an $\alpha = 0.05$.

6. Summary

The Bonferroni correction was not used because Bonferroni corrections are employed to reduce Type I errors (i.e., rejecting H_0 when H_0 is true, or false positives) when multiple tests or comparisons are conducted (in our case the null hypothesis is that there is no difference in the evaluation metrics between the groups) and it is used to check if any of the metrics abide to the hypothesis.

In our case, and in order to launch the experiment, the evaluation metrics must throw a significant difference but not only on one or some of the metrics but all of them must show that significant difference (in other words, we want to see differences, otherwise the experiment would fail or no effect observed when introducing the changes).

Although the Bonferroni method can always be used, we find that the Bonferroni method is not appropriate in this context. We could probably consider other methods, like the Holm-Bonferroni method, which is just as general but less conservative, or just ignore the whole issue just like we did in our experiment.

Particularly in our experiment, we found that the difference in the Gross conversion between groups to be statistically and practically significant @ 95% Confidence Interval, therefore rejecting the null hypothesis (they are different). On the other hand, the Net conversion was found to be neither statistically nor practically significant @ 95% Confidence Interval.

7. Recommendation

Bottom line, the business case was to test if the introduced changes would produce the following effects:

- improve the overall student experience and
- improve coaches' capacity to support students who are likely to complete the course
- all of that without decreasing the revenue.

The launch criteria should measure the two main goals of the experiment:

- Decrease the number of students who sign up for the free trial but don't continue past the free trial,
- without decreasing the revenue.

In terms of metrics, to launch the experiment, we must observe a practically and statistically significant decrease in gross conversion as well as a practically and statistically significant not decrease for net conversion and retention.

With the analysis performed we **cannot recommend the launch of the experiment** without carrying out other complementary studies because:

- The Gross conversion is both statistical and practically significant; that is good because the results could be that less students unlikely to convert to the full course would sign-up in trials.
- On the other hand, and because the Net conversion is neither statistically nor practically significant and because the confidence interval includes negative numbers, there is the risk that we could end up with less revenue because less students would convert (as net figure) is a possibility that we cannot rule out.

It could be also true that the financial savings derived from the Gross conversion results outperform the possible loss of revenue from the Net conversion, resulting a net financial gain....it could also be possible that it happens the other way round (more likely, in our opinion). At this point we do not have exact financial figures in order to get a final conclusion.

The student (client) satisfaction is something that will have to be measured with other methods like surveys. This is a very important topic because a satisfied customer is a good ambassador of the product, moreover in today's "connected" world, where people freely express their experiences in the web (Facebook, forums...etc.).

For all that reasons we recommend the design of other experiments before committing to the real implementation of the new features.

FOLLOW-UP EXPERIMENT

Now, we would like to design some experiment to reduce the number of frustrated students who cancel early in the course.

First of all, I would conduct some marketing research about the different causes that could trigger an early drop in the course, in the form of in-depth interviews with recent drop-off students, regular surveys at the cancellation time, secondary research on the same topic from different sources, expert judgement advice...etc.

Some causes would involve students just “looking around”, “the course is not what I thought”, “I have some programming skills but a lot more is required”, lack of real interest...among others.

We could implement a better explanation of the courses so the students really know what topics are exactly covered, to which depth, which programming, statistics, mathematics skills are needed in order to succeed the course in a timely manner leaving the student with a sensation of fulfillment. That could be done with a 30 min talk with one of the coaches; the coach could evaluate with some questions the suitability of the tandem student knowledge/ambition – course difficulty/achievement.

Apart from the time commitment (the current experiment) and a better explanation of the courses with the coach, I would add a skill testing exam in the technologies to be used with the possibility of being redirected to other courses, if needed (for instance, redirection from the machine learning course to a Statistics or Python course). This exam could be waived if the coach issues a favorable report about the prospective student; in doubt, the student is redirected to the exam.

The flow would be as follows:

Visitor (cookie) → sign up for free trial (user-id). Now two options: continue with free trial or pre-enrollment coach interview / level test exam → full enrolled student (payments).

So, the experiment would set up a control and an experiment group randomly chosen among the visitors who signed up for the trial period.

The experiment group will have to go through a talk/interview with the coach and then pass a skill level test (if required by coach) while the control group will not be directed to a coach talk plus exam, but just continue with the regular free trial, towards enrollment (if any).

Null hypothesis would be that the implementation of these new features will not decrease early cancellations by a statistically and practically significant amount.

The **Unit of Diversion** would be the user-id; the users have already signed up for the trial so we can trace them along the whole process.

The **Invariant** metric would be also the user-id. It is the unit of diversion, besides, the experiment starts before the visitor (now, signed in) decides to go for an interview with the coach.

The **Evaluation** metric would be the Retention. Some calculations will have to be made to find out if that metric is statistically and practically significant. If the significance tests are passed, that would indicate good chances

to have found out differences between the experiment and the control group. Apart from different figures, we are expecting to find the student retention to increase.

Should these variables behave in the desired way, we could then proceed to extend this experimental program to all the website's visitors. Otherwise some other experiments would have to be designed or additional studies to be carried out.

References

- Bonferroni Method: https://en.wikipedia.org/wiki/Bonferroni_correction
- Bonferroni adjustments: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112991/>
- Stackexchange: <http://stats.stackexchange.com/search?q=ab+testing>
- AB Testing Plan: <http://conversionxl.com/how-to-build-a-strong-ab-testing-plan-that-gets-results/>
- 3 AB real examples: <http://blog.hubspot.com/marketing/a-b-testing-experiments-examples>