Design an A/B test: final project

Xin Pang

Experiment Design

Figure 1 shows the process flow starting from when a user look at course description till enrolling in a trial or only accessing free course materials.

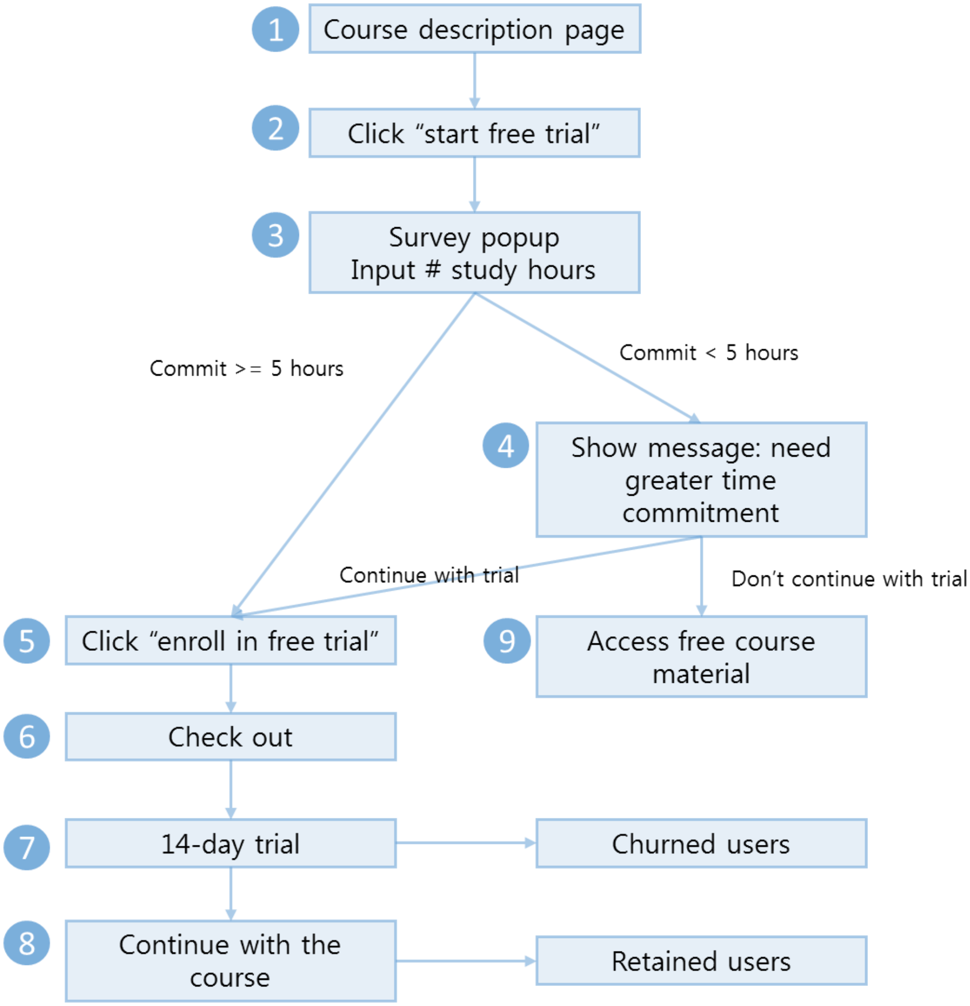
1 Course description page
2 Click "start free trial"
3 Survey popup
Input # study hours
Commit >= 5 hours
Commit < 5 hours
4 Show message: need greater time commitment
Continue with trial
Don't continue with trial
5 Click "enroll in free trial"
9 Access free course material
6 Check out
7 14-day trial
Churned users
8 Continue with the course
Retained users

Figure 1. Experiment process flow

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

Table 1. Metric choices

Metrics	**Invariant**	**Evaluation**	**Reason**
Number of cookies: That is, number of unique cookies to view the course overview page. (d_{min}=3000)	Yes	No	As shown in Figure 1, the main purpose is to measure whether adding step 4 will reduce churned users while keeping the amount of retained users at a comparable level. Calculating the number of cookies in the course overview page should be part of the sanity check, since course overview happens before enrollment to free trial, and we want to make sure that about the same amount of cookies assigned to each group.
Number of user-ids: That is, number of users who enroll in the free trial. (d_{min}=50)	No	No	This metric cannot be used as an invariant, because it happens after the free trial screener. ~~Also, it cannot be used as an evaluation metric. The goal is to measure whether screener has an effect on the free trial and paid lessons or not. Increased number of users who choose to enroll cannot assure that an increased number of users will stay till finishing the trial and courses.~~ It can be used as an evaluation metric because one of the goals is to reduce the number of less committed users (users who cannot commit at least 5 hours a week). However it is not an ideal one since a comparison between absolute numbers can result in wrong conclusion. Given that gross/net conversion and retention are already chosen, this one is no longer needed.
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)	Yes	No	Because this happens before the free trial screener, it could also be included as invariant metrics, to make sure number of people who clicked the button is about the same, such that number of users who see the screener and who do not is about the same.
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)	Yes	No	If this rate is very different for the control and experiment group, it is likely that users recruited in both groups have different characteristics. For example, if there are more students in control group and more working people in experiment group, then it is likely that less users click the “start free trial” button in the control group (could be due to financial reasons etc.). Therefore, even if the number of users who clicked the start free trial button are about the same, due to different characteristics of both groups, the result can be very different. Hence the click-through-probability should be included as an invariant.
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)	No	Yes	After applying the screener, it can happen that the number of enrolled users will decrease due to lack of commitment for the required number of hours. Since part of our experiment is to “reduce the number frustrated users upfront” before the free trial, this metric is relevant as an evaluation metric.
Retention: That is, number of	No	Yes	Ideally this should be chosen as evaluation metric

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)			since it is linked to the goal that adding the screener should not significantly reduce the number of students to continue past the free trial. In fact, if adding the screener is successful, the retention rate should be increased or at least about the same as before. However, as mentioned later in the section “sizing”, in order to assure enough power for the experiment, the number of page views required is very impractical, thus in reality this is not chosen.
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)	No	Yes	This is chosen as evaluation metric since it is linked to the ultimate goal of increasing the number of paid and committed students.

Overall, following is an indication of this experiment to be successful, based on the selected evaluation metrics:

- Gross conversion: decreased because of filtering out students who cannot commit >5 hours.
- Net conversion: be the same or increased because of getting committed students to enroll and finish the course.
- Retention: be increased because of less frustrated students in the checkout and trial period.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

Table 2. Measuring standard deviation

	Population	Sample	Standard error
Unique cookies to view page per day:	40000	5000	
Unique cookies to click "Start free trial" per day:	3200	5000 * 0.08 = 400	
Enrollments per day:	660	5000 * (660/40000) = 82.5	
Click-through-probability on "Start free trial":	0.08		
Probability of enrolling, given click:	0.20625		Sqrt(0.20625 * (1-0.20625) / 400) = 0.020230604
Probability of payment, given enroll:	0.53		Sqrt(0.53 * (1-0.53) / 82.5) = 0.054949012
Probability of payment, given click	0.1093125		Sqrt(0.1093125 * (1-0.1093125) / 400) = 0.015601545

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

- Gross conversion, net conversion: expected to be about the same.
- Retention: empirical variability can be different from the analytic estimate.
- Empirical estimate is required when the unit of diversion and unit of analysis is different. The reason of different variability is that, most distributions are based on the assumption of independence, choosing different units will make this assumption no longer valid. Since there can be a lot of correlation, a difference of analytic and empirical variability is expected.
- In this experiment, the unit of diversion is cookie. For gross conversion and net conversion, the unit of analysis is cookie as well. Thus the analytic and empirical would be comparable. For retention, the unit of diversion is cookie, but the unit of analysis is user ID, since they are different, we would expect some difference.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

I choose not to use Bonferroni correction. ~~There are 3 metrics for the experiment, if Bonferroni correction is to be applied, then the significance level for each will be 0.05/3 = 0.016667. This requires a much larger sample of users in order to run the test. In our case, the population is only ~40000. Hence the test can be too conservative such that no test will be passed.~~ The idea of using Bonferroni correction is to reduce the false positives when tracking multiple metrics at the same time, when the launch decision is made based on one of them. The goal for the experiment is to 1) reduce the number of frustrated students who left the free trial AND 2) without significantly reducing the number of students to continue past the free trial and eventually complete the course. Since both needs to be satisfied, this is not the situation where Bonferroni will help.

Using the calculator on this website (http://www.evanmiller.org/ab-testing/sample-size.html), we get results shown in Table 3. For the probability of payment given enroll (retention), it is clear that the required number of page views is exceeding the possible page views too much, thus resulting in very long test duration. Thus running the experiment will be very expensive. Also, if the experiment is running too long, there will be more unpredictable factors in the experiment settings, making the result inaccurate to be interpreted. Therefore, I only choose the gross and net conversion as the evaluation metrics.

Given the above, a larger number should be chosen as the minimum sample size of page views, in order to provide enough power to the experiment for both metrics. Thus the number of page views is 685325.

Table 3. Calculating number of page views

Metrics	Population	Min size per group	Min size two groups	Min sample size pageviews
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	25835	51670	51670 / 0.08 = 645875
Probability of payment, given enroll:	0.53	39087	78174	78174 / (660 / 40000) = 4737818
Probability of payment, given click	0.1093125	27413	54826	54826 / 0.08 = **685325**

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

Given that the trial period is 14 days, we need to make sure that at least some users go through the full cycle such that the result won't be influenced by halfway users too much. Also, considering the potential different user behaviors in weekdays and weekends, the number of days would be a multiple of 7. The choice of number of days for running the experiment is 28, and based on this, the fraction of traffic is 685325 / 28 / 40000 = 0.61189732.

Risk evaluation: from the user side, the students will not get any negative influence in case of being chosen in the experiment group. The only extra step is to input the number of study hours and based on own selection decide to continue or not. It is up to the user to decide, and there is no financial / physical / psychological effect on this. Also, there is no privacy information violation regarding this, given that the number of hours is not personally identifiable information. From audacity's side, theoretically there would not be any significant negative influence on the company's finance as well as the brand image. There could be slight loss of revenue given that users who are not fully committed will go pass the trial period and continue with payment if they were not asked with the study hours. However, if they first continue and then drop out soon, it is also not a good use of the mentor and learning resources.
Overall, since the risk level is low, we can proceed with this experiment.

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)

Table 4. Summary stats control and experiment groups

	Control	Experiment
Pageviews total	345543	344660
Clicks total	28378	28325
Click-through-probability	0.08212581	0.08218244

Table 5 and 6 show the calculation for page view check and clicks respectively. The idea is that given the probability that a user will be randomly assigned to control / experiment group, 0.5 in the ideal case, by calculating the confidence interval, will the observed fraction of users assigned to the controlled group be falling into the range. By calculating the CI, the observed value falls into the range, thus both checks pass.

Table 5. Sanity checks page views

Number of cookies	
Probability	0.5
Standard error	Sqrt (0.5 * 0.5 / (345543+344660)) = 0.000601841
Confidence interval upper	0.5 + 1.96 * 0.000601841 = 0.501179608
Confidence interval lower	0.5 - 1.96 * 0.000601841 = 0.498820392
Observed	345543 / (345543+344660) = 0.500639667

Table 6. Sanity checks clicks

Number of clicks	
Probability	0.5
Standard error	Sqrt (0.5 * 0.5 / (28378+28325)) = 0.002099747
Confidence interval upper	0.5 + 1.96 * 0.002099747 = 0.504115504
Confidence interval lower	0.5 - 1.96 * 0.002099747 = 0.495884496
Observed	28378 / (28378+28325) = 0.500467347

Table 7 shows the calculation for click through probability. By calculating the click through probability from the controlled group, and its confidence interval, we can compare and see whether the click through probability in the experiment group falls in the range. Since it is the case, this check also passes.

Table 7. Sanity checks click through probability

Click through probability	
Probability of click	28378 / 345543 = 0.082125814
Standard error	Sqrt (0.08212581 * (1-0.08212581) / 345543) = 0.000467068
Confidence interval upper	0.08212581 + 1.96 * 0.000467068 = 0.083041267
Confidence interval lower	0.08212581 - 1.96 * 0.000467068 = 0.081210360
Observed	0.082182441

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)

Gross conversion and net conversion are selected metrics.
For gross conversion, hypothesis H0 is the gross conversion rates for both control and experiment group are about the same. Calculation are shown in table 8. Since the final CI is [-0.029123, -0.011986], which is not including 0 and is larger than the d_{min} = 0.01, the result is statistically significant and practically significant.

Table 8. Effect size test – gross conversion

Gross conversion	
d-min	0.01
alpha	0.05
Enroll-control	3785
Click-control	17293
Enroll-experiment	3423
Click-experiment	17260
Pooled probability	(3785+3423) / (17293+17260) = 0.208607
Pooled SE	Sqrt(0.208607 * (1-0.208607) * ((1/17293) + (1/17260)) = 0.004372
Gross conversion control	3785/17293 = 0.218875
Gross conversion experiment	3423/17260 = 0.198320
Margin error	0.004372 * 1.96 = 0.008568
d-hat	0.198320 – 0.218875 = -0.020555
CI lower	-0.020555 – 0.008568 = -0.029123
CI upper	-0.020555 + 0.008568 = -0.011986

For net conversion, hypothesis H0 is the net conversion rates for both control and experiment group are about the same. Calculation are shown in table 9. Since the final CI is [-0.011604, 0.001857], which is including 0, the result is neither statistically significant nor practically significant.

Table 9. Effect test size – net conversion

Net conversion

d-min	0.0075
alpha	0.05
Enroll-control	2033
Click-control	17293
Enroll-experiment	1945
Click-experiment	17260
Pooled probability	(2033+1945) / (17293+17260) = 0.115127
Pooled SE	Sqrt(0.115127 * (1-0.115127) * ((1/17293) + (1/17260)) = 0.003434
Gross conversion control	2033 / 17293 = 0.117562
Gross conversion experiment	1945 / 17260 = 0.112688
Margin error	0.003434 * 1.96 = 0.006730
d-hat	0.112688 – 0.117562 = -0.004873
CI lower	-0.004873 - 0.006730 = -0.011604
CI upper	-0.004873 + 0.006730 = 0.001857

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

In order to perform the sign test, gross and net conversion for both the control and experiment groups are calculated on a per day basis. Also, the difference between control and experiment group is calculated, as shown in table 10.

For gross conversion, the main goal is to decrease the rate, since we want to reduce the number of users who can potentially be frustrated due to lack of time commitment. The hypothesis here is H0 there is no difference between control and experiment group over the gross conversion rate. Based on this, there are 19 out of 23 successful outcomes (marked as red). Using the calculation tool in this link (http://graphpad.com/quickcalcs/binomial1/), the p value is 0.0026. Here the two tailed p value is used, because the test is about whether there is difference or not, instead of how many more successful outcomes there should be. Given that 0.05 is the significance level and 0.0026 < 0.05, we can draw the conclusion that the result is statistically significant.

Same applies to the net conversion rate. There are 10 successful outcomes (marked as green), because the desired outcome should be an increase value in the experiment group compared with the controlled group. After calculating the two tailed p value, it is 0.6776, which is not statistically significant.

Table 10. Sign test

Date	GrossCon	GrossExp	GrossDiff	NetCon	NetExp	NetDiff
Sat, Oct 11	0.195051	0.153061	-0.041990	0.101892	0.049563	-0.052330

Sun, Oct 12	0.188703	0.147771	-0.040933	0.089859	0.115924	0.026065
Mon, Oct 13	0.183718	0.164027	-0.019691	0.104510	0.089367	-0.015144
Tue, Oct 14	0.186603	0.166868	-0.019735	0.125598	0.111245	-0.014353
Wed, Oct 15	0.194743	0.168269	-0.026474	0.076464	0.112981	0.036517
Thu, Oct 16	0.167679	0.163706	-0.003974	0.099635	0.077411	-0.022224
Fri, Oct 17	0.195187	0.162821	-0.032367	0.101604	0.056410	-0.045194
Sat, Oct 18	0.174051	0.144172	-0.029879	0.110759	0.095092	-0.015667
Sun, Oct 19	0.189580	0.172166	-0.017414	0.086831	0.110473	0.023643
Mon, Oct 20	0.191638	0.177907	-0.013731	0.112660	0.113953	0.001294
Tue, Oct 21	0.226067	0.165509	-0.060558	0.121107	0.082176	-0.038931
Wed, Oct 22	0.193317	0.159800	-0.033517	0.109785	0.087391	-0.022394
Thu, Oct 23	0.190977	0.190031	-0.000946	0.084211	0.105919	0.021708
Fri, Oct 24	0.326895	0.278336	-0.048559	0.181278	0.134864	-0.046414
Sat, Oct 25	0.254703	0.189836	-0.064868	0.185239	0.121076	-0.064163
Sun, Oct 26	0.227401	0.220779	-0.006622	0.146893	0.145743	-0.001150
Mon, Oct 27	0.306983	0.276265	-0.030718	0.163373	0.154345	-0.009028
Tue, Oct 28	0.209239	0.220109	0.010870	0.123641	0.163043	0.039402
Wed, Oct 29	0.265223	0.276479	0.011255	0.116373	0.132050	0.015676
Thu, Oct 30	0.227520	0.284341	0.056820	0.102180	0.092033	-0.010147
Fri, Oct 31	0.246459	0.252078	0.005619	0.143059	0.170360	0.027301
Sat, Nov 1	0.229075	0.204317	-0.024758	0.136564	0.143885	0.007321
Sun, Nov 2	0.297258	0.251381	-0.045877	0.096681	0.142265	0.045584

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

Bonferroni correction is not used. The intention for using Bonferroni correction is to reduce the false positives when tracking multiple metrics at the same time and the launch decision is made based on one of them. The goal for the experiment is to 1) reduce the number of frustrated students who left the free trial AND 2) without significantly reducing the number of students to continue past the free trial and eventually complete the course. Especially the second goal is important because we don’t want to lose paid users. In our result it is already shown that the test result for the second sub-goal is not statistically significant. Thus there is no need to use Bonferroni correction.

The conclusion from the sign test is the same as that of the effect size test, i.e. gross conversion result is statistically significant, but not the net conversion.

Recommendation

Make a recommendation and briefly describe your reasoning.

The screener works well in reducing the number of people to continue from click to enrollment. ~~But it does not work well in keeping the comparable amount of users that will continue and past the 14-day trial.~~ The confidence interval for net conversion rate is [-0.011604, 0.001857], although it includes 0 and the positive side, given the predefined level is [-0.0075, 0.0075], the calculated CI does include the negative range [-0.0075, 0] and even further. This means the net conversion rate might go down to an extent beyond the business expectation. Based on this, we should not launch the change.

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.

Considering the goal is to make sure that students commit enough time in the study, and frustration level not only depend on the time to be committed, it might also result from the money-time devotion. For a working person they might be satisfied to study 3 hours a week while keep learning for a longer time. While for a student with no income yet they might find it important to devote as much time as possible in order to finish all the lessons. With this in mind, another screener might be added, asking e.g. how many weeks would you like to spend if to finish a course. For this the net and gross conversion would still be the valid evaluation metrics. The invariant metrics and unit of diversion stay the same.

Experiment design description

An experiment can be designed in the middle of the free trial period: by end of first week trial, we can send users an email, asking that if they want to directly go with the subscribing in the program, in return they will unlock some study content that is uniquely available in Audacity and not in the normal course settings. This will help in differentiating students who are really enthusiastic about the course from less committed students.

The hypothesis is that for highly committed users, since their motivation is already strong, providing the possibility of unlocking some study material will increase their willingness to enroll and pay. For less committed users, it will not make significant difference since they are not strongly motivated.

In this experiment, the invariant metric would be number of users who start free trial, since all the steps happen after the trial period starts.

Considering this, the suitable metrics would be the following:

- Percentage of users who enrolled and paid via discount by end of first week: p1
- Percentage of users who enrolled and paid by end of second week: p2
- Overall percentage of users who enrolled and paid in the two weeks: p-total

A successful indicator for the experiment would be: p1 > p2, and $\text{p-total}_{\text{experiment}} > \text{p-total}_{\text{control}}$

This is because it does not make much difference for motivated users whether it is the first week or the second week to enroll, if they are anyways committed to finish the course. Therefore they would be more willing to enroll with the additional benefits. After the first week since committed users are already enrolled, we would observe a decrease in the percentage of users who enroll and pay, since we have less motivated users in the user base.
This extra benefit would also help to encourage uses who are otherwise not motivated enough to enroll, thus we would observe an increased overall percentage compared with control group.

Unit of diversion: given that all the steps happen after users entering the trial period, the unit of diversion would be user ID.