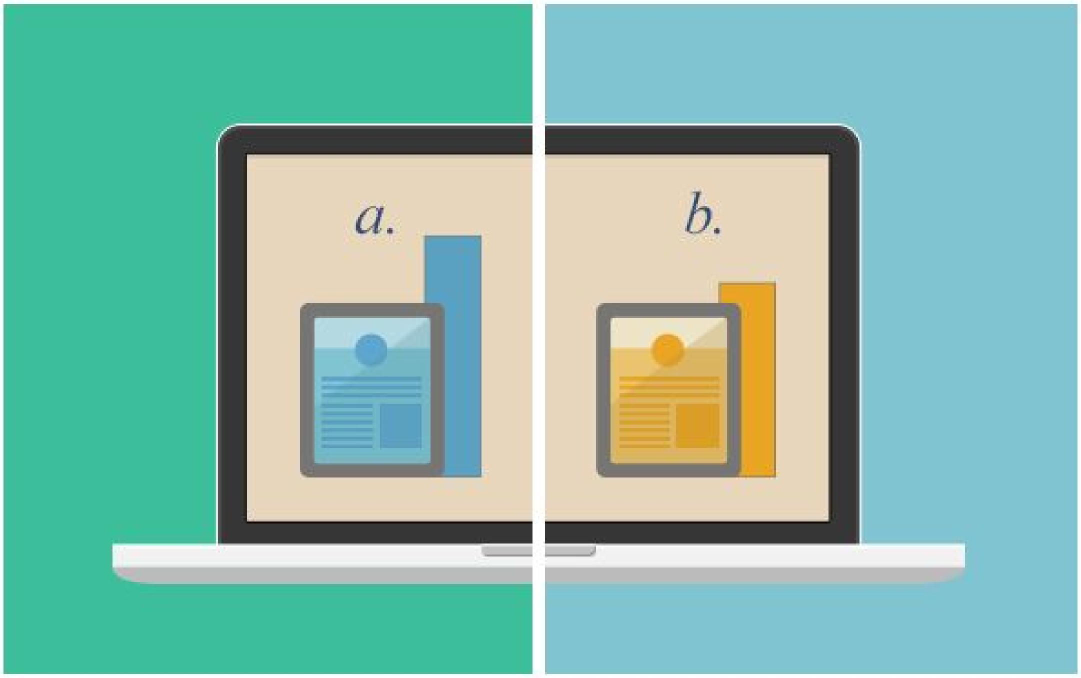
Udacity Free Trial Screener A/B Test

Andrew Ehsaei, 03/27/2017



Experiment 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

The proposed change is to add a pop-up question to users who click the "start free trial" button and ask their predicted time to devote to the course. If the student responds with less than 5 hours per week, the user is informed that the course requires a greater time commitment.

Null Hypothesis (Ho)

The change will not affect the number of students that remain enrolled through the free trial period and will not affect the number of students that remain enrolled beyond the free trial period.

Alternative Hypothesis (Ha)

The change will reduce the number of students that quit before the free trial period ends without significantly reducing the number of students that remain enrolled beyond the free trial period.

Metric Choice

In order to measure changes between the experiment and constant, we must choose a set of metrics. There are two types of metrics: invariant and evaluation metrics. Invariant metrics are variables that will remain unchanged between the control and experiment. Evaluation metrics are variables that will change between the control and experiment and allow us to measure differences between the groups.

Metric	**Type**	**Definition**
Number of Cookies	Invariant	Number of unique cookies to view the course overview page.
Number of User-Ids	-	Number of users who enroll in the free trial.
Number of Clicks	Invariant	Number of unique cookies to click the "start free trial" button (which happens before the free trial screener is triggered).
Click-Through Probability	Invariant	Number of unique cookies to click the "start free trial" button divided by the number of unique cookies to view the course overview page.
Gross Conversion	Evaluation	Number of user-ids to complete checkout and enroll in the free trial divided by the number of unique cookies to click the "Start free trial" button.
Retention	Evaluation	Number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of user-ids to complete checkout.
Net Conversion	Evaluation	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.

Number of Cookies

This metric measures the number of unique cookies to view the course overview page and should not change based on the experiment, therefore it is an invariant and not an evaluation metric. Since the change to the website will occur at an event after users have already viewed the course overview page, this metric should remain consistent between the control and experiment groups.

Number of User-ids

This metric measures the number of users who enroll in the free trial. We are testing a change to the website that would affect this metric directly and we expect a difference in this variable between the control and experiment groups. However, this metric is a raw count of the number of enrollments and is not normalized. This raw count can vary widely depending on the number of cookies or clicks in a given day and between the control and experiment groups. We have a metric called the gross conversion, which will inform impact on enrollments that we can normalize. The number of user-ids will not be used as a metric.

Number of Clicks

This metric measures the number of unique cookies to click the "start free trial" button. The stipulation on this metric is that the click happens before the free trial screener is triggered, aka our experiment event. Since this metric measures events that occur before the experiment begins, it should remain consistent between the control and experiment groups and be classified as an invariant metric.

Click-through Probability

This variable measures the number of unique cookies to click the "start free trial" button divided by the number of unique cookies to view the course overview page. Both the click event and the view event both occur before the experiment screener event, this metric will remain consistent between the control and experiment. Therefore, this metric is an invariant metric and not an evaluation metric.

Gross Conversion

The next metric in the list is the "gross conversion." This metric is a ratio of the number of enrollments divided by the number of "start free trial" clicks. This metric will possibly be affected by our experiment, since the change will present the users with additional information before enrolling in the free trial. This metric will measure a difference between the control and experiment and therefore is an evaluation metric.

Retention

The next metric is the "retention" metric. This variable is another ratio of enrollments beyond the 14-day boundary divided by the number of user-ids to complete checkout. The enrollments beyond the 14-day boundary and number of users to complete checkout are both values that the experiment is attempting to affect. This ratio will show a difference between the control and experiment and therefore is an evaluation metric.

Net Conversion

The last metric is the "net conversion" metric. This variable is another ratio of the number of user-ids to remain enrolled past the 14-day boundary divided by the number of unique cookies to click the "start free trial" button. The number of user-ids to remain enrolled past the trial period will show a difference between the control and experiment groups, therefore "net conversion" is an evaluation metric. The number of cookies to click the button should be consistent between the control and experiment, but the numerator of the ratio will differ.

Launch Criteria

The goals of the experiment are (1) to reduce enrollments by unprepared students (2) without significantly reducing the number of students who complete the free trial and make at least one payment. To achieve these goals we require that the evaluation metrics show specific differences between our experiment and control.

Gross conversion is the ratio of trial enrollments to trial clicks. The first goal requires this ratio to decrease since we will see a reduction in enrollments by unprepared students. Retention is the ratio of post-trial enrollments to user-ids. The first goal will require that this ratio increases in the experiment over the control since there will be a reduction in the number of unprepared students who were less likely to enroll. Net conversion is the ratio of post-trial enrollments to trial clicks. The second goal requires the net conversion not to decrease between the experiment and control groups. The number of students who complete the trial and make at least one payment should not be significantly reduced.

Measuring Standard Deviation

For each evaluation metric, I made an analytic estimate of its standard deviation, given a sample size of 5000 cookies visiting the course overview page. Here is some basic website data to use in these calculations:

Variable	**Value**
Sample Size	5000
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 samples of the same size (**N**) are repeatedly randomly drawn from a population, and the proportion of successes in each sample is recorded (**p^**), the distribution of the sample proportions (i.e., the sampling distribution) can be approximated by a normal distribution given that **(N) × (p^) > 5**. The standard deviation (**SD**) of the distribution of sample proportions has the following relationship:
SD = sqrt[**p^** × (1 - **p^**) / **N**]

Gross Conversion Standard Deviation

p^ = Probability of enrolling, given click
p^ = 0.20625

N = Sample Size × (Proportion of viewers to click "Start free trial")
N = Sample Size × **(** Clicks / Pageviews)
N = 5,000 × (3,200 / 40,000)
N = 400
SD = sqrt[**p^** × (1 - **p^**) / **N**]
SD = sqrt[0.20625 × (1 - 0.20625) / 400]
SD = 0.0202

Retention Standard Deviation

p^ = Probability of payment, given enroll
p^ = 0.53
N = Sample Size × (Proportion of viewers to enroll)
N = Sample Size × **(** Enrollments / Pageviews)
N = 5,000 × (660 / 40,000)
N = 82.5
SD = sqrt[**p^** × (1 - **p^**) / **N**]
SD = sqrt[0.53 × (1 - 0.53) / 82.5]
SD = 0.0549

Net Conversion Standard Deviation

p^ = Probability of payment, given click
p^ = 0.1093125
N = Sample Size × (Proportion of viewers to click "Start free trial")
N = Sample Size × **(** Clicks / Pageviews)
N = 5,000 × (3,200 / 40,000)
N = 400
SD = sqrt[**p^** × (1 - **p^**) / **N**]
SD = sqrt[0.1093125 × (1 - 0.1093125) / 400
SD = 0.0156

In this experiment, the unit of diversion is the "cookie." It is the subject that we use for testing and comparison empirically. In the calculations of gross conversion and net conversion, the analytic estimate of the standard deviation was also done in cookie units, so I expect the empirical variability to be comparable to the result. In the Retention calculation, the units used in this calculation were enrollments and not cookies, therefore I expect the analytic estimate and the empirical variability to be different.

Sizing

Number of Samples vs. Power

In order to calculate the number of samples needed to power the experiment, we must discuss the two error types. A type I error occurs when the null hypothesis (H0) is true, but is rejected. It is asserting something that is absent, a false hit. A type I error may be likened to a so-called false positive (a result that indicates that a given condition is present when it actually is not present).

The type I error rate or significance level is the probability of rejecting the null hypothesis given that it is true and is denoted by the Greek letter α (alpha) and is also called the alpha level. Often, the significance level is set to 0.05 (5%), implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis.

A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected. It is failing to assert what is present, a miss. A type II error may be compared with a so-called false negative (where an actual 'hit' was disregarded by the test and seen as a 'miss') in a test checking for a single condition with a definitive result of true or false. A Type II error is committed when we fail to believe a truth. The rate of the type II error is denoted by the Greek letter β (beta) and related to the power of a test (which equals 1−β).

To calculate the number of pageviews needed for our experiment, we will use a significance level of 0.05 and a power level of 80%. I'll use this online tool to calculate each sample size.
α = 0.05
B = 0.2

Gross Conversion Pageviews

Baseline conversion rate = 20.625% (p^)
Minimum Detectable Effect = 1%
Sample Size = 25,835 (clicks)
Pageviews = 25,835 (clicks) x 12.5 (ratio of pageviews / click) x 2 (experiment and control)
Pageviews = 645,875

Retention Pageviews

Baseline conversion rate = 0.53 (p^)
Minimum Detectable Effect = 1%
Sample Size = 39,087 (enrollements)
Pageviews = 39,087 (enrollments) x 60.606 (pageviews / enrollment) x 2 (experiment and control)
Pageviews = 4,737,818.182
Note: The amount of pageviews needed here is very large and not a feasible number for the experiment. I'll drop this metric from the experiment.

Net Conversion Pageviews

Baseline conversion rate = 0.1093125 (p^)
Minimum Detectable Effect = 0.75%
Sample Size = 27,413 (clicks)
Pageviews = 27,413 (enrollments) x 12.5 (pageviews / click) x 2 (experiment and control)
Pageviews = 685,325
This is the largest page view size calculated (dropping the retention metric), so this will be the number of pageviews needed for both the Gross Conversion and Net Conversion metrics.

We have a situation here where we are tracking multiple metrics, and therefore we are more likely to see a statistically significant result by chance. We have a choice of two methods to avoid this pitfall, the first option is to assume the metrics are independent, and the other is to use the Bonferroni correction. The problem with the Bonferroni correction is that this method may be overly conservative. I'll go with method 1 and forgo the Bonferroni correction.

Duration vs. Exposure

The first thing to consider when deciding the duration and exposure of the experiment is the risk involved for the users undergoing the experiment. Since this experiment does not deal with sensitive user account information or put anyone at risk of harm, the risk is low. Since the risk is low, the client may want the shortest possible duration to complete the experiment. The more traffic that is diverted, the shorter duration required to complete the experiment. Since the screener experiment is a low risk I would divert all the traffic to the experiment and estimate the shortest duration possible.

Number of pageviews needed: 685, 325 pageviews
Pageviews per day: 40,000 pageviews / day
Fraction of traffic to divert: 1
Duration: 685,325 / (40,000 × 1)
Duration: 18 days

Experiment Analysis

Sanity Checks

The experiment was conducted and the results populated in a spreadsheet. For each of the invariant metrics, we can calculate a confidence interval for the expected value and the observed value and see if the metric passes a sanity check.

Variable	**Description / Formula**
Z score (Z)	1.96 (for 95% confidence interval)
Probability (p^)	0.5 (probability of success for 2 outcomes)
Standard Error (SE)	sqrt[p^ × (1 - p^) / Ntotal]
Margin of Error (m)	Z × SE
Confidence interval (CI)	[p^ - m , p^ + m]
Observed	Ncontrol / Ntotal

Number of Cookies Confidence Interval

Probability (p^) = 0.5 (cookie / no-cookie)
N = 345,543 (Control) + 344,660 (Experiment)
N = 690,203
SE = sqrt[0.5 × 0.5 / 690,203]
m = 1.96 × 0.00060184
m = 0.0011796
Confidence interval: [0.4988, 0.5012]
Observed: 0.5006
Pass Sanity Check (Observed lies within confidence interval)

Number of Clicks Confidence Interval

p^ = 0.5 (click / no-click)
N = 28,378 (Control) + 28,325 (Experiment)
N = 56,703
SE = sqrt[0.5 × 0.5 / 56,703]
m = 1.96 × 0.0020997

m = 0.0041155
Confidence interval: [0.4959, 0.5041]
Observed: 0.5005
Pass Sanity Check (Observed lies within confidence interval)

Variable	**Description / Formula**
Pooled Probability (p^pool)	(Xcontrol + Xexperiment) / (Ncontrol + Nexperiment)
Pooled Standard Error (SEpool)	sqrt[p^pool × (1 - p^pool) × (1/Ncontrol + 1/Nexperiment)
Difference (d^)	Xexperiment / Nexperiment - Xcontrol / Ncontrol
Confidence interval (CI)	[d^ - m , d^ + m]

Click-through Probability on "Start free trial" Confidence Interval

p^pool = (28,378 + 28,325) / (345,543 + 344,660)
p^pool = 0.08216
SEpool = sqrt[0.08216 × (1 - 0.08216) × (1/ 28,378 + 1/28,325)]
SEpool = 0.00066106
m^ = 1.96 × 0.00066106
m = 0.0012956776
d^ = 28,325/344,660 - 28,378/345,543
d^ = 0.000056627
Confidence interval: [-0.0013, 0.0013]
Observed: 0.0001
Pass Sanity Check (Observed lies within confidence interval)

After calculating each of the invariant metric confidence intervals and the observed probability values, each of these metrics pass the sanity check. For the cookies and number of clicks calculations I used the standard error and confidence interval formulas for strict counts. For the click-through-probability calculation I used the pooled standard error and difference equations, since the sample groups were not the same size.

Result Analysis

Effect Size Tests

For each of the evaluation metrics, we can compute a confidence interval around the difference between the experiment and control groups. This will allow us to indicate whether each metric is statistically and / or practically significant. I've decided to not use the Bonferroni correction with these calculations.

Gross Conversion Confidence Interval

Gross Control = enrollments / clicks
d^ = (3,423 / 17,260) - (3,785 / 17,293)
d^ = -0.02055
p^pool = (3,423 + 3,785) / (17,260 + 17,293)
p^pool = 0.208607
SEpool = 0.00437
m = 0.0085685
Confidence interval: [-0.0291, -0.012]
d^ > m, statistically significant

d^ > dmin value of 2%, practically significant

Net Conversion Confidence Interval

Net Conversion = payments / clicks
d^ = (1,945 / 17,260) - (2,033 / 17,293)
d^ = -0.0048737
p^pool = (1,945 + 2,033) / (17,260 + 17,293)
p^pool = 0.115127
SEpool = 0.0034341
m = 0.0067309
Confidence interval: [-0.0116, 0.0018]
d^ < m, Not statistically significant
d^ < dmin value of 2%, Not practically significant

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.

For the Gross Conversion metric, the confidence interval includes 0, and the difference is larger than the margin of error and therefore statistically significant. The difference calculated is greater than the 2% minimum significance boundary, so the result is practically significant as well.

For Net Conversion, the confidence interval does not include 0, and the difference is smaller than the margin of error and therefore not statistically significant. The difference calculated is less than a 2% minimum significance level and therefore not practically significant either.

Sign Tests

The sign test is a statistical method to test for consistent differences between pairs of observations, such as the weight of subjects before and after treatment. Given pairs of observations (such as weight pre- and post-treatment) for each subject, the sign test determines if one member of the pair (such as pre-treatment) tends to be greater than (or less than) the other member of the pair (such as post-treatment).

For each of the evaluation metrics, I'll do a sign test using the day-by-day data, and report the p-value of the sign test. I'll also report whether the result is statistically significant. I'll use the QuickCalcs online sign test to perform the calculation.

For Gross Conversion, we have 4/23 successes where the number of enrollments were larger in the experiment over the control. **This results in a two-tail P value of 0.0026 and given our significance level α = 0.05, this result is statistically significant.**

For Net Conversion, we have 10/23 successes where the number of payments were larger in the experiment over the control. **This results in a two-tail P value of 0.6776, which is larger than our significance level of 0.05. Therefore this result is not statistically significant.**

Summary

In summary we can review the results of our evaluation metrics. In the effect size tests, the gross conversion metric results were statistically and practically significant. The statistical significance was confirmed using the sign test. The tests show that the change to the website did cause the gross conversion to drop by roughly 2% in the experiment over the control. This means that users were less likely to enroll in the experiment.

The net conversion metric results were not statistically or practically significant. The effect size tests showed a small difference between the experiment and control, but was found to be neither statistically and practically significant. The sign test confirmed that the difference was not statistically significant as well. The take-away from this result is that the website change did not show to increase the number of enrollments beyond the 14-day trial period.

To reiterate the null hypothesis: “The change will not affect the number of students that remain enrolled through the free trial period and will not affect the number of students that remain enrolled beyond the free trial period.” The first part of the null hypothesis was disproven with the gross conversion metric, however the second part was not. The change did not affect the number of enrollments beyond the 14-day trial period and therefore I cannot reject the null hypothesis.

When performing multiple comparisons, the chance of a statistical error increases with the number of metrics. The Bonferroni correction reduces the likelihood of seeing a statistically significant result by chance (decreased false positives) at the expense of power (increased false negatives).□ The choice of whether or not to use the correction should be based on how our launch decision would be impacted by statistical error. This in turn is related to our launch criteria.□ False positives have the greatest impact when any of the metrics satisfied can trigger launch, since a single false positive will govern the decision.□ False negatives have the greatest impact when all metrics must be satisfied to trigger launch, since a single false negative can govern the decision. In our experiment, our launch criteria requires that all metrics must be satisfied to trigger the launch, so false negatives have the greatest impact. Since using the Bonferroni correction would increase the likelihood of false negatives, I have chosen not to use it.

Recommendation

The requirement for the gross conversion metric was to show a decrease in the experiment over the control and this was confirmed within our results. The gross conversion did show a decrease in the experiment over the control and the result was tested to be both statistically and practically significant. This launch criteria requirement was met by the experiment.

The second requirement was that the net conversion should not show a decrease between the experiment and control. This requirement was not met by our experiment. The result showed a decrease in the experiment group over the control group, but failed to achieve statistical or practical significance. Furthermore, the confidence interval of the net conversion included the negative of the practical significance boundary. That is, it's possible that this number went down by an amount that would matter to the business.

The experiment failed to meet our second requirement of our launch criteria. We failed to achieve statistical and practical significance for net conversion. The results show there is a risk that the website change could result in losing users who remain enrolled past the 14-day boundary. It is for these reasons that I recommend not proceeding with the website change.

Follow-Up Experiment

The goal of the experiment is to find a way to increase the number of students that remain enrolled beyond the 14-day trial period. I would propose a change to the website to encourage potential students to begin and increase time spent on the program. I would incentivise the potential students with a discounted enrollment fee if a percentage of progress has been made.

My hypothesis would be that by incentivising student progress during the free trial period in the form of discounted enrollment fees, the number of enrollments beyond the 14-day period would increase. The details of the offer could be a $100 credit towards enrollment if you complete the first 20 lessons by the end of the trial-period.

The experiment will be done on users after they have enrolled in the free-trial, so we will need metrics that measure differences within this period. We could use the number of users who enroll in the trial period as an invariant metric. For an evaluation metric we could use a ratio of the number of users to remain enrolled beyond the 14-day trial period over the number of users to enroll in the free trial. This evaluation metric would be a great way to measure the number of users who cancel within the 14-day trial period within the experiment and control. Since we will be dealing with enrolled users, the subject or unit of diversion would be the user-id.

Resources

https://en.wikipedia.org/wiki/A/B_testing
https://www.optimizely.com/ab-testing/
https://www.r-bloggers.com/standard-deviation-vs-standard-error/
https://onlinecourses.science.psu.edu/stat200/node/43
https://en.wikipedia.org/wiki/Type_I_and_type_II_errors
http://www.evanmiller.org/ab-testing/sample-size.html
http://graphpad.com/quickcalcs/binomial1.cfm
https://en.wikipedia.org/wiki/Sign_test