Experiment Design

Metric Choice

In the experiment, the following metrics can be measured:

- **Number of cookies** (d_{min}=3000)
- **Number of user-ids** (d_{min}=50)
- **Number of clicks** (d_{min}=240)
- **Click-through-probability** (d_{min}=0.01)
- **Gross conversion** (d_{min}= 0.01)
- **Retention** (d_{min}=0.01)
- **Net conversion** (d_{min}= 0.0075)

Chosen invariant metrics (these should not change between experiment and control):
1. Pageviews.
2. Clicks (when a user clicks to start a free trial).
3. Clickthrough: clicks divided by pageviews.

We need all of these metrics to not be significantly different between experiment and control datasets. Otherwise the evaluation metrics will be affected too much by the variance in these parameters.

The following were chosen as evaluation metrics:
1. Gross conversion.
This should reduce as an effect of this experiment. The practical significance boundary would be passed if the gross conversion reduces by 0.01 (1%). Therefore we should set up the experiment as a one-way statistical test where we reject the null hypothesis on the negative side. And we accept the test if we accept the alternate hypothesis.

2. Retention.
This should increase as an effect of this experiment. The practical significance boundary would be passed if the retention increases by over 0.01 (1%). Therefore we should set up the experiment as a one-way statistical test where we reject the null hypothesis on the positive side. And we accept the test if we accept the alternate hypothesis.

3. Net conversion.
This should not change as an effect of this experiment. Therefore we would want the null hypothesis to hold where the practical significance boundary should not be crossed on either direction (the boundary is 0.0075 or 0.75%). Therefore we accept the test if we accept the null hypothesis.

In order to pass the experiment, we want all of the metrics to pass. This is because the initial goal of the experiment had several criteria: First, we wanted to reduce the amount of people who start the free trial and proceed to enroll (to reduce the amount of people who end up realising too late that they do not have the motivation/time to finish the course). Second, we want the rate of people who start the free trial and proceed to pay to remain the same (we don't want to discourage users who are willing to proceed with the course). Third, we want the rate of enrolled users who proceed to pay to increase.

Measuring Standard Deviation

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

For each evaluation metric, an analytical estimate for their variability was made using rough baseline data of 5000 pageviews:

Metric	N	p	std
Gross conversion	400	0.20625	0.0202
Retention	82.5	0.53	0.0549
Net conversion	400	0.1093125	0.0156

The units of analysis (N) are quite low for each metric, but especially for Retention. Therefore, I would not completely trust these values, but especially so for the Retention-metric. If possible, making an empirical estimate for the variance of Retention would be a good idea. The fact that the standard deviation of retention is significantly higher than the rest supports this notion.

Sizing

Number of Samples vs. Power

Bonferroni correction was chosen to not be used in the analysis. This is because the hypothesis for each evaluation metric are considered to be separate (independent) and all evaluation metrics are necessary to pass in order to make the final decision.

Therefore, the parameters for the decision for the required sample size are the following:

alpha: 0.05
beta: 0.2

Metric	Conversion rate (p)	Significance boundaries (dmin)	Units of analysis needed	Pageviews needed (in total for both Experiment and Control)
Gross conversion	0.20625	0.01	25835	645875
Retention	0.53	0.01	39115	4741212
Net conversion	0.1093125	0.0075	27413	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.)

Since this experiment is a very low risk one, it doesn't seem to be necessary to divert any data. Therefore the experiment data can be collected faster without causing any real risk.

The parameters for the duration and exposure are hence:

traffic direction ratio: 1.0
daily pageviews: 40000

Therefore, the days needed to run the experiment in order to get enough data is: 119 days.

It is evident that the days needed to run the experiment in order to get enough data for Retention is massive. Even though it is a very promising metric for this experiment, gathering enough data for it to be usable takes just way too long. Therefore it is necessary to omit Retention from the experiment.

The experiment is still possible to run without the Retention-metric, since the Gross conversion and Net conversion together implicitly also measure the Retention metric. But is absolutely necessary that both of these metrics pass in order to proceed with the experiment.

Without Retention as a metric, the days needed to run experiment is only 18 days.

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.)

Sanity checking is used to verify the invariant metrics in the experiment- and control-datasets. It should be expected that the invariant metric distributions do not differ significantly between experiment and control datasets.

These are the chosen invariant metrics:
1. Pageviews.
2. Clicks (when a user clicks start a free trial).
3. Clickthrough: clicks divided by pageviews.

The sanity check hypothesis are the following:
1. The page views should not differ significantly between experiment and control.
2. The clicks should not differ significantly between experiment and control.
3. The clickthrough of control should not differ significantly from the clickthrough of experiment.

Then, a confidence interval around the p-values of the sanity check are defined, and then we check if the observed values fall within that confidence interval.

The table below shows the results of the sanity checks.

Invariant metric	p-value	CI	observed value	passes?
Pageviews	0.5	0.4988, 0.5012	0.5006	Yes
Clicks	0.5	0.4959, 0.5041	0.5005	Yes
Clickthrough	0.0821	0.0812, 0.0830	0.0822	Yes

As seen in the table, the sanity tests pass for each invariant metric.

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.)

The effective size analysis can be seen in the table below.

Evaluation metric	dhat	N [Control, Experiment]	SE	m	CI	Statistical significance	Practical significance
Gross conversion	-0.0206	[17293, 17260]	0.0044	±0.0086	-0.0291, -0.0120	Yes	Yes
Net conversion	-0.0049	[17293, 17260]	0.0034	±0.0067	-0.0116, 0.0019	No	No

When reviewing the initial hypothesis, it was defined that the alternate hypothesis for a Gross conversion test is desirable and is accepted if the experiment gross conversion is significantly less than the gross conversion in control (in layman's terms, significantly less people who click "start free trial" will go on and enroll). The table defines with both statistical and practical significance that this is true.

For Net Conversion, on the other hand, we do not want the value to change significantly between control and experiment (the same number of people who click "start free trial" will go on and make a payment on the course). Hence, in this case we want the null hypothesis to hold. The table shows that the null hypothesis will hold with both statistical and practical significance.

Therefore, both tests will pass the effect size analysis.

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.)

To verify the experiment results, we will do a sign test for the day-by-day data to see whether the differences of control and experiment will be correctly signed on most days within the experiment. Therefore we can be certain that the experiment result is supported by the entire dataset and not just a few outliers within the data.

The results of the sign test can be seen in the table below. The p-values were calculated using an online calculator [2].

Metric	# of days	Positives	p-value
Gross conversion	23	4	0.0026
Net conversion	23	10	0.6776

The results of the sign-test indicate that for the Gross conversion-case we are likely to reject the null hypothesis, and the direction of the day-by-day data (only 4 are positive) indicate that the direction of the data is skewed towards the negative side.

The net conversion has such a high p-value that the null hypothesis can not be rejected, which is exactly what we were looking for.

Summary

The evaluation metrics used for the experiment were Gross conversion and Net conversion. Initially, Retention was used as well but it was deemed to require too much time to gather enough data to create a viable experiment using that metric.

Because the evaluation metric tests were deemed to be independent and the experiment itself absolutely required all tests to pass in order to proceed (especially after dropping the Retention-metric), the use of Bonferroni correction was deemed unnecessary.

In the effective size tests, both metrics passed with both statistical and practical significance. It is therefore possible to infer from the experiment that it is likely that the amount of users on trial who will enroll for courses will be reduced without affecting the amount of users on trial who will make an initial payment.

The sign tests also passed, indicating the we should accept the alternate hypothesis for Gross conversion and accept the null hypothesis for Net conversion. The sign test design for Gross conversion could've been improved however. It could've been a one-way test to indicate whether the sign is negative in most of the day-by-day data. Now it was a two-way test to indicate whether the sign differs towards the positive or negative direction. It could, however, be argued that the current experiment already tested that, although more implicitly.

Recommendation

Because all evaluation metrics pass the tests with both statistical and practical significance, the recommendation is to proceed with the change.

Follow-Up Experiment

An overall goal of the nano degree enrollments should be to increase the amount of users who end up compelled to start the free trial. Then a larger amount of users would move on within the funnel to the enrollment- and payment-stages.

Currently the amount of free trial is tiny (1 week) compared to the amount of commitment the nano degrees requires (around 6 months to 1 year). If the free trial is increased, does the amount of people who enroll increase?

Therefore, the experiment hypothesis would be:
Given an increased free trial phase, do the amount of people who enroll increase?

Since we want to not only increase people who "click free trial", but in the end want to increase the enrollment numbers, this should be an evaluation metric.

What we would like to evaluate is two-fold. First of all, we would like to know if the amount of people who try the free trial increase if we increase the free trial period. Second, we want to know if this also increases the amount of people who enroll to the course.

It also should be noted that this change is likely not going to increase the rates of which people enroll after starting a free trial, but are likely increase the clickthrough rates on page viewers starting the free trial.

A definate invariant metric that is needed should be that the amount of people who view the course pages remains standard.

To summarize the experiment:

Hypothesis:
Given an increased free trial phase, do the amount of people who enroll increase?

Evaluation metrics:
Clickthrough on “Start free trial” from pageviews .
Enrollment rate from pageviews.

Invariant metrics:
Pageviews

Unit of diversion:
Pageviews (cookies)

References:

[1] http://www.evanmiller.org/ab-testing/sample-size.html
[2] http://graphpad.com/quickcalcs/binomial1.cfm