Experiment Design

Metric Choice

Metric name	Reason
1. Number of cookies (to view the course page)	Should not change, because at this point, experiment has not started yet. Chosen as invariant. Not a good choice for an evaluation metric, since it happens before experiment starts.
2. Number of user IDs (to enroll in free trial)	Should not change significantly, but I did not select this as invariant, because there is a possibility of a change. Not a good choice for an evaluation metric either, since it's not a ratio that can be normalized for different in size between control and experiment groups.
3. Number of clicks (on "Start free trial")	Also should not change, if everything works as it should, and chosen as invariant. Same reason as #1 for not selecting it as evaluation metric.
4. Click-through probability (on "Start free trial")	This metric is the ratio of metrics #1 and #3. If they don't change, this shouldn't either. Chosen as invariant. Same reason as #1 for not selecting it as evaluation metric.
5. Gross conversion (N-enrolled / N-clicked)	Expect this to decrease, so can't be invariant. Selected as evaluation metric.
6. Retention (N-paid / N-enrolled)	Initially, I selected it as evaluation metric. But it required too many days of running the experiment to see a significant change. So I went back and deselected it. Can't be invariant, because change is expected.
7. Net conversion (N-paid / N-clicked)	Experiment would be a success if this metric does not decrease. I selected this as an evaluation metric, since it's necessary for it to work together with Gross conversion.

We're looking for two things in this experiment:

1. To have less students that left free trial without completing the course because they didn't have time.
2. But not reduce the number of students that continue after free trial to complete the course.

This means we want to increase Retention, but not reduce Net conversion. Also we expect Gross conversion to decrease.

Measuring Standard Deviation

Metric name	SD
Gross conversion	0.0202
Retention	0.0549
Net conversion	0.0156

Analytical variability can be used when unit of diversion matches unit of analysis. Here is unit of analysis for our metrics:

Metric name	Unit of diversion	Unit of analysis	Matches?
Gross conversion (N-enrolled / N-clicked)	Cookie	Cookie	Yes
Net conversion (N-paid / N-clicked)	Cookie	Cookie	Yes

In this case, analytical variability would probably match empirical variability.

Sizing

Number of Samples vs. Power

If not using Bonferroni correction, I need the following number of samples:

Metric name	SD
Gross conversion	645875
Retention	4741212
Net conversion	685325

Duration vs. Exposure

I need the following number of days, depending on fraction of traffic and subset of metrics:

Metric name	Diverting 50%	Diverting 75%	Diverting 100%
Gross conversion	33	22	17
Retention	238	159	119
Net conversion	35	23	18

I think the experiment is not risky, because:

- People would not be hurt during the experiment.
- No sensitive data is collected.

However:

- It's possible that new implementation may have a bug.
- It's possible that a technical problem may arise when deploying the new version.

For example:

- the client-side javascript was tested in latest versions of IE, Firefox, Safari and Chrome, but fails in older versions.
- we have a new website behind the load balancer, and that's where we redirect experiment traffic. But the new website is not configured correctly to access database, so students can't really enroll.

Normally, I would divert 50% of the traffic, just to make sure we don't accidentally make a mistake that affects too many students. However, we have a requirement that experiment should not last over 30 days. So I'll go with diverting 75%. That'll require 23 days to run.

Experiment Analysis

Sanity Checks

Metric name	Lower bound	Upper bound	Observed	Pass
Number of cookies	0.4988	0.5012	0.5006	Yes

Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through probability	-0.0013	0.0013	0.0001	Yes

Result Analysis

Effect Size Tests

Metric name	Lower	Upper	Practical significance boundary	Statistical significance	Practical significance
Gross conversion	-0.0291	-0.0120	0.01	Yes	Yes
Net conversion	-0.0116	0.0019	0.0075	No	No

For gross conversion, results are:

- statistically significant, because the CI does not include 0.
- practically significant, because the CI does not include the given boundary of 0.01.

For net conversion, results are:

- not statistically significant, because the CI does include 0.
- not practically significant.

Sign Tests

Metric name	p-value	Statistical significance
Gross conversion	0.0026	Yes
Net conversion	0.6776	No

Summary

For this experiment, both metrics should change in the expected way: Gross conversion should decrease, and Net conversion stay the same (or about). Bonferroni correction is supposed to be used for independent metrics, to reduce the chance of Type 1 errors (False Positive), at the expense of increasing the chance of Type 2 errors (False Negative). Here, metrics are dependent, so the experiment already has higher chance of FN and lower chance of FP. I did not use the correction.

Recommendation

I would not recommend launching the change, because I could not confirm statistical significance of change in both evaluation metrics. Moreover, confidence interval of Net conversion includes negative of the practical significance boundary, so the results may actually be harmful to the business.

Follow-Up Experiment

On loading any website page, user would receive a persistent cookie (say we name it “experiment2”) with a value that contains a unique string ID (GUID). That cookie would be set to expire many years in the future (which is not really “never” but certainly longer that duration of experiment). Users, of course, can always clear or not allow cookies; it’s an edge case and we can’t guarantee consistent experience for these users. Otherwise, using a unique ID value of “experiment2” cookie, user would be randomly included in either control group or experiment group. Once user is logged in, he would stay in the same group, but now tracked by user ID instead of cookie, to provide consistent experience.

For students included in experiment group, on every course overview page, I would display additional information:

1. Percent of students that completed the class, out of the total enrolled.
2. Average number of weeks or months it took those students to complete the class.
3. Average number of hours per week that each successful student spent on this class.

This follows the same logic as the first experiment: we want to set the student’s expectations.

- 1st would expose the level of difficulty of this class, so that the future student could see right away what his chances are.
- 2nd would point out the long-term commitment needed.
- 3rd would point out every day’s commitment expected.

I would select the metrics as follows:

Metric name	**Reason**
1. Number of cookies to view the course page	Invariant, because this metric is measured before experiment starts.
2. Total number of clicks on both “Start free trial” and “Access course for free”	Evaluation. Expect this to not change significantly.

3. Retention	Evaluation. Expect to increase.

Hypothesis is that we can:

- Increase retention.
- But not decrease total number of users that take the course, in its free form or not.

Unit of diversion would be a cookie, because I want consistent experience for a user, whether he is logged in or not.