Steven Yeh

P7: Design an A/B Test

Experiment Design

Metric Choice

Invariant Metrics

For invariant metrics, I decided to use **number of cookies** (unique cookies to view course overview page), **number of clicks** (number of users who click to start the free trial), and the **click-through-probability** (number of clicks / number of cookies).

Within the scope of the experiment, the **number of cookies** should be the same in both groups since this occurs before users click to start the free trial. In addition, **number of cookies** is used as the unit of diversion and as such should be split evenly between the groups. **Number of clicks** should be the same between the control and experiment groups since nothing before the click was changed or affected by the experiment. Since **click-through-probability** is a function of the previous two invariant metrics, **click-through-probability** should also be an adequate invariant metric.

Number of user-ids (number of users enrolling in the free trial) is not an invariant metric because changing how users proceed through the free trial process is expected, or at least hopefully supposed to have an effect on this metric.

Similar logic can be applied to **gross conversion, retention,** and **net conversion**; each of these metrics is a function of at least one variable that occurs chronologically after where the control and experiment groups diverge. **Gross conversion** is partially calculated by the number of user-ids, and as such cannot be invariant. User-ids who have made payments, which occur 14 days after enrollment, affect **retention** and **net conversion** and cannot be invariant since these values are expected to differ between the control and experiment groups.

Evaluation Metrics

Invariant metrics by definition cannot be used as evaluation metrics because they should stay constant between groups. Therefore, **number of cookies, number of clicks**, and **click-through-probability** were not chosen as evaluation metrics**.**

Number of user-ids was not chosen as an evaluation metric because it was not a best choice for this experiment. Practically speaking, randomly assigning a large amount of subjects in both groups is not guaranteed, and differences in sample number between the two groups could negatively impact the results and skew our conclusions. In addition, there

are rate metrics that control for this possible discrepancy between group sizes such as net conversion or gross conversion.

Retention (the number of user-ids that made payments divided by the number enrolled) initially looks like a suitable evaluation metric because it is reasonable to guess that retention might be higher in the experiment group, as users enroll with a time commitment in mind. However, when calculating the duration of the experiment, including **retention** as an evaluation metric increases the duration of the experiment from around a month to over half a year! Subsequently, the required page views increases substantially from 685325 to well over a million. Considering how unreasonable a timeframe this experiment would potentially be if **retention** as included, it would be wise to omit it as an evaluation metric.

Gross conversion (number of user-ids to enroll divided by number of clicks) was chosen as an evaluation metric because the hypothesis addresses users continuing to the free trial. If the number of users that enroll decreases as a result of the experiment, than the gross conversion should also decrease. Users who realize that they do not have the time commitment could be deterred from enrolling.

Net conversion (number of user-ids to make payments divided by number of clicks) was chosen as an evaluation metric because the hypothesis also addresses keeping the number of students who continue past the free trial (thus making payments) similar between the control and experiment groups. Therefore, it is imperative to evaluate the proportion of users continuing past the free trial and examine if there is no difference in this metric between the control and experiment groups. Notice that both evaluation metrics control for group size, as they are rates.

To launch the experiment, both the proposed reduction in **gross conversion** and change in **net conversion** need to be examined. In the scope of this experiment, if the point estimate and the confidence interval of the **gross conversion** metric are below the practical significance boundary, then the experimental change is statistically significant. Since the **net conversion** is hypothesized not to change, a statistically significant change is not necessary in order to recommend a launch. Instead, one should ascertain that there is no negative effect on **net conversion** with the experimental group, as stated by the hypothesis. In other words, the statistical outcomes for both metrics should be as expected in order to recommend an experiment launch. One expects to see a statistically significant difference in gross conversion and no significant change in net conversion.

Steven Yeh

Measuring Standard Deviation

The analytic estimate of the standard deviation for **gross retention** was calculated to be **0.0202**, given a sample size of 5000 cookies. The analytic estimate of the standard deviation for **net retention** was calculated to be **0.0156**. Since the unit of diversion is a **cookie**, and the unit of analysis for both evaluation metrics is the number of unique **cookies** that start the free trial, the two units are the same and the analytical estimate is likely to match the empirical variability. Therefore, the analytical estimate can be used for both evaluation metrics. In addition, samples are most likely independent when the units of diversion and analysis match, which fulfills the independent sampling assumption.

Sizing

Number of Samples vs. Power

I decided not to use the Bonferroni correction. Using the evaluation metrics: **gross conversion** and **net conversion**, and an alpha = 0.05 and beta = 0.2, the number of pageviews needed to appropriately execute the experiment was calculated to be **685325** pageviews.

Duration vs. Exposure

I decided to divert **0.75** of Udacity's traffic to the experiment. As Udacity receives 40,000 unique cookies viewing the page per day, the days needed to run an experiment of 40,000 * 0.75 = **30,000** pageviews per day was calculated to take (685325 views) / (30,000 views/day) = 22.8442 rounded up to **23 days**.

I chose a fraction of 0.75 to lower the experiment duration to a few weeks. In general, it is risky to test 100% of users to an experimental change as it is unclear how the reception to this change will be. However, financial risk is directly monitored in this experiment, so the decline would be spotted very quickly. It is also generally unwise to subject every user to an experimental change that has not been finalized. On the other hand, reducing the fraction to 0.25 would increase the length of the experiment to 70 days, which is not practical for client expectations. A 0.75 diversion fraction preserves 25% of the traffic from an experimental change while still maintaining the experiment duration at a realistic timeframe of 23 days, or slightly over 3 weeks. The experiment itself should not be considered to be a high risk since cookie-based diversion is anonymous and user consent is not an issue. Therefore, the relatively higher diversion is not a risky choice since there is no collection of sensitive data or possibility of harm.

Steven Yeh

Experiment Analysis

Sanity Checks

Here are the 95% confidence intervals for my invariant metrics.

Invariant Metric	Lower Bound	Upper Bound	Observed Value	Passes Check
Number of cookies	0.4988	0.5011	0.5006	Yes
Number of clicks on "Start free trial"	0.4959	0.5041	0.5005	Yes
Click-through-probability on "Start free trial"	0.0812	0.0830	0.0822	Yes

As all invariant metrics passed the sanity check (the observed value is within the 95% confidence interval), we can proceed to the rest of the analysis.

Result Analysis

Effect Size Tests

Here are the 95% confidence intervals around the difference between the experiment and control groups.

Evaluation Metric	Lower Bound	Upper Bound
Gross conversion	-0.0291	-0.0120
Net conversion	-0.0116	0.0019

Gross conversion is **statistically significant** because its 95% confidence interval does not include 0.0. Gross conversion is also **practically significant** because its 95% confidence interval excludes the practical significance boundary specified for this metric on the negative side, -0.01. We expect the experiment to generate a negative change beyond the practical significant value in the negative direction.

Net conversion is **not statistically significant** because its 95% confidence interval contains 0.0. Net conversion is also **not practically significant** because its 95% confidence interval contains the practical significance boundary on the negative side (-0.0075). Therefore, it is very possible that the changed net conversion rate calculated from the experiment does not exceed the practical significance.

Sign Tests

Here are the sign test results for each of the evaluation metrics.

Evaluation Metric	p-value	Statistical Significance
Gross conversion	0.0026	Yes
Net conversion	0.6776	No

Gross conversion has a statistically significant p-value because 0.0026 is less than the 0.05 significance level. Conversely, net conversion does not have a statistically significant value since its p-value is much higher than the 0.05 significance level.

Summary

I chose not to use the Bonferroni correction because both the net conversion and gross conversion are relevant as specified by the hypothesis. Because there were specific expectations for both metrics (negative gross conversion, unchanged net conversion), and both effects are examined within this experiment, a Bonferroni correction is unnecessary as using one might hinder the experiment outcome and cause the conclusions to be overly conservative.

The effect size hypothesis tests and sign tests results were as expected with no discrepancies between the two tests. The negative change in gross conversion was statistically and practically significant. The change in net conversion was neither statistically nor practically significant.

Recommendation

The gross conversion reduction is practically significant **and** the net conversion is expected to not to decrease significantly. Both evaluation metrics behave accordingly with their expected effects as shown by both the effect size and sign test results. In addition, all the invariant metrics also pass the sanity checks. However, there is evidence that the net conversion may actually decrease, which would not satisfy the goal of not reducing revenue. The confidence interval for net conversion is [-0.0116, 0.0019] while the negative of its practical significance boundary is d_{min} = -0.0075. Since the CI includes the practical significance boundary, the possibility of a decrease in net conversion during the experiment exists. The day-by-day data revealed an actual negative difference of -0.0049, but this result is not sufficient to draw any definite conclusions. I would recommend **not launching and further testing**, as a true negative change in net conversion is possible. It would also be good to investigate a "payments" metric, or the number of user-ids that remain enrolled past 14 days as it affects net conversion.

Steven Yeh

Follow-Up Experiment

If the goal were to reduce the number of frustrated students who cancel early in the course, I would experiment with a 50% tuition return for students who complete the course. This notification would be displayed as a pop-up message on top of the course overview page before the user clicks to "Start free trial." The experiment and control groups would be determined by assigning unique (per day) cookies to each group. Cookies assigned to the experiment group would then be subject to the pop-up message, while cookies assigned to the control group would not see the intervention and proceed as normal.

The **hypothesis** would be that a 50% tuition return would increase the number of students to continue past the free trial since this is a lucrative deal, while reducing the number of frustrated students who cancel early since they are now incentivized to finish, possibly preferring a certificate obtained at half the cost instead of refunding with no certificate.

Similar to the main experiment, I would choose to measure **gross conversion** and **net conversion** since they are resistant to group sizes. If **retention** and **click-through-probability** could be measured without significantly increasing the amount of pageviews and the experiment time, then I would also choose these metrics for their resistances to group size. **The click-through-probability, retention,** and **net conversion** would be tested for an increase since number of clicks, number of user-ids, and number of payments should increase, while **gross-conversion** would be tested for no change since the experiment occurs before users click "Start free trial."

The **unit of diversion** would still be a **cookie** since most of the evaluation metrics are calculated by number of cookies in the denominator. Since **retention** is determined by number of user-ids, the empirical estimate of the variability should be used for this metric.

In this experiment, only **number of cookies** could be an invariant metric. In both control and experiment groups, the number of unique cookies viewing the course overview page should be the same. **Number of clicks** and **number of user-ids** cannot be invariant metrics because hopefully these amounts increase in the experiment group. In addition, I would not use these two metrics as evaluation metrics because randomly assigning a large amount of subjects between groups is not guaranteed to be equal. Differences in these groups could potentially skew these metrics since they depend on count.