Design an A/B Test

By William Autry

1 Experiment Design

Make design decisions for an A/B test, including which metrics to measure and how long the test should be run. Analyze the results of an A/B test that was run by Udacity and recommend whether or not to launch the change.

1.1 Metric Choice

The metrics below were considered for this experiment:

- **Number of cookies:** That is, number of unique cookies to view the course overview page.
- **Number of user-ids**: That is, number of users who enroll in the free trial.
- **Number of clicks**: That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is triggered).
- **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.
- **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.
- **Retention**: That is, number of 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.
- **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.

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

Invariant Metrics:

1) **Number of cookies** – Cookies are assigned to both the experimental and control group so this value should not vary. Would not be a good evaluation metric because it's not affected by the experimental change.
2) **Number of clicks** – This value should also be identical, since a visitor has to click the "start free trial" button prior to the experiment. Would not be a good evaluation metric because it's not affected by the experimental change.
3) **Click-through-probability** – This value is 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. So this too should be constant since the button is clicked prior to the experiment. Would not be a good evaluation metric because it's not affected by the experimental change.

Evaluation Metrics:

1) **Gross conversion** – This metric will show whether costs will decrease or not after the screener is utilized. The metric would not be a good choice for our invariant metrics, because the number of visitors enrolled in the program is a value that is directly affected by the experimental change.
2) **Net conversion** – This metric will show how changes affect revenues. It would be idea for this experiment to not reduce this metric, if it does decrease the change should be abandoned. This would not be a good invariant metric because the number of visitors that remain enrolled after 14 day trial is affected by the experimental change.

Expected Results: The experiment can be launched if Gross conversion significantly decreases and there is no significant change in Net conversion. It is important that there is no decrease in the Net conversion metric, because a decrease could foretell a decrease in revenue. It is equally important to consider the confidence intervals and the risk factor prior to launching the experiment.

Unused Metrics:

- **Number of user-ids** – This metric would not be a good choice for our invariant metrics, because the number of users who enroll in the free trial is dependent on the experiment. It is not the best metric as it is not normalized (so Net Conversion is definitively better), but it could potentially be an evaluation metric. The number of user IDs is usable as an evaluation metric because it would track the first part of the hypothesis; namely whether we will reduce the number of students to continue past the free trial.
- **Retention** – This metric is not a good invariant choice because the number of user-ids to remain enrolled is dependent on the experiment. This metric could be used as an evaluation metric since it provides insight into the results of the experiment. However, when I tried including it as an evaluation metric I removed it after reaching the sizing section. As an evaluation metric, this metric increased the time period needed to collect the desired number of pageviews.

1.2 Measuring Standard Deviation

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.

Evaluation Metric	Standard Deviation
Gross Conversion	0.0202
Net conversion	0.0156

The Unit of Diversion for Gross Conversion & Net Conversion is the number of cookies in their denominator, which is also equal to the Unit of Analysis. So the analytical estimate should be comparable to the empirical variability.

1.3 SIZING

1.3.1 Number of Samples vs. Power

Bonferroni is not relevant to what we are trying to discover.

Using $\alpha = 0.05$ & $\beta = 0.2$, I calculated that I will need 685,325 pageviews.

1.3.2 Duration vs. Exposure

I would divert 100% of Udacity's traffic to the experiment. That's 40,000 experiment pageviews per day, at that rate it would take 18 days to gather the required number of pageviews.

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

I initially tried 40%, but it extended the duration of the experiment to 43 days which is not idea. This experiment is low risk since there is only one extra question added to an existing process. Also the nature of the data is not sensitive information.

2 EXPERIMENT ANALYSIS

2.1 SANITY CHECKS

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

Invariant Metric	Lower Bound	Upper Bound	Observed	Passes
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through-probability	0.0812	0.0830	0.0822	Yes

2.2 RESULT ANALYSIS

2.2.1 Effect Size Tests

For each of the Evaluation Metrics, I computed a confidence interval around the difference.

Evaluation Metric	Lower Bound	Upper Bound	Statistical Significance	Practical Significance
Gross Conversion	-0.0291	-0.0119	Yes	Yes
Net Conversion	-0.0116	0.0018	No	No

2.2.2 Sign Tests

For each of the Evaluation Metrics, I ran a sign test using the day-by-day data.

Evaluation Metric	p-value	Statistical Significance
Gross Conversion	0.0026	Yes

Net Conversion	0.6776	No

2.2.3 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.

I chose not to use Bonferroni correction, since the method is used to decrease Type I false positives at the expense of increasing Type II errors. This experiment requires all of the metrics meet the acceptance criteria; Significant Decrease in Gross Conversion and no decrease in Net Conversion. This increases the vulnerability of the experiment to Type II errors. The results from the Effect Size & Sign Tests show that Gross Conversion will statistically and practically decrease, while Net Conversion is not affected.

2.3 RECOMMENDATION

Make a recommendation and briefly describe your reasoning.

Gross Conversion was discovered to be negative, and is Statistically & Practically significant, which translates to a lowered operating cost due to the discouragement of trial signups that are unlikely to convert. Net Conversion was found to be Statistically & Practically insignificant. However, the confidence interval of the net conversion does include the negative of the practical significance boundary. This means that it is possible that the number went down by an amount that would matter to Udacity.

So, I would recommend that Udacity find an alternative method to screen users or abandon the idea completely. The use of the screener as currently designed would negatively impact signups and could affect revenue.

3 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.

To create a follow-up experiment investigating the same principle issue of student retention, of course split students evenly and randomly into each the control & experimental group. I would allow the newly signed up student to attempt the course work for a week uninterrupted. However, once a week of the two week trial has passed, check-in on the student to poll what Udacity could do to help them succeed now that they have been introduced to the curriculum.

Hypothesis: If Udacity introduces the user to all available resources, they may be more inclined to seek assistance and continue through any course they may find challenging.

Metrics:

- **Number of user-ids** – This metric can be used as the invariant metric since it will not be affected by the experiment.
- **Retention** – This metric can be used as the evaluation metric since it is directly correlated to the core of the proposed experiment.

Unit of Diversion: Number of user-ids could be used for this since we are evaluating users that have already enrolled and reached the midway point of their trial.

Only implement the change investigated in this experiment if retention is positive & statistically/practically significant.

4 REFERENCES

I. Bonferroni Correction – https://en.wikipedia.org/wiki/Bonferroni_correction
II. A/B Testing – https://vwo.com/ab-testing/