Subject: A/B Online Testing Memo
To: Interested Parties - SIP partners
From: Daniel Gonzales, Lena Tom, Analyst Institute
Date: May 11, 2017

Introduction

Conventional wisdom for different messages and modes only take an organization so far. Analyst Institute has seen that even top performing poll tested messages rarely performs in rank order when used in a member ask or a petition. By learning how to run basic testing A/B testing and analytics, an organization can calculate post per action in all of their programs thereby optimizing each campaign. Additionally, organizations have the potential to develop reports that captures this sort of data from multiple campaigns enabling the organization to compare performance by issue, mode, and message. This will serve to provide important localized expected performance and give context specifically for how different reproductive rights and abortion out loud messages perform compared to each other as well as help sketch out which digital modes are most appropriate for which audiences.

Research Questions

The design process for A/B testing can help address a series of research questions regarding the efficacy of an organization's online outreach. This may include testing for the most effective email subject lines, sign up pages, action rates on petitions, social media share rates and more.

However, A/B testing may not be for all organizations. Testing requires a fairly large list of targets to conduct tests, since the response rates to online actions are often very small (less than 1%). If your membership list is under 50,000 you are unlikely to be able to conduct robust A/B tests. But even with a small list, some groups may be able to run occasional A/B tests on tactics of great importance. If you have a small list, you may want to consult with someone familiar with A/B testing before determining the viability of a test.

A/B Testing Planning

The following is a recommendation document for A/B testing[1].

Taking advantage of every opportunity: View every piece of communication as an opportunity to test something. Creating an assumption that a test will be included in the process will help smooth the testing process. The tests may be small - such as running segments of lists against each other with different subject lines. Each testing opportunity will allow you to build on your body of knowledge.

[1] Special thanks to Amelia Showalter of A/B Cafe & Pantheon Analytics for these insights

Brainstorming & Deciding on priorities: The opportunities for A/B tests can be limitless - starting with small, easy to agree upon tests will help you set a baseline understanding of testing as well as the activity on your list. As your testing ability matures, you may prioritize tests that might produce a "multiplier" effect. For example, is there a test that can reactivate your inactive users, thereby increasing your program's impact?

Learning from failure: Building a culture that acknowledges that many of our tests will produce no results (or worse results) is as important as the testing itself. Often times, testing tells us that our guts instincts are off track - focusing on lessons learned and applying them to future tests and program will improve that instinct.

Sharpening your hypothesis: A/B testing is a way to compare distinct tactics and measure effect, but it is not typically a way to answer large programmatic questions. For example, we could test two emails, one using Abortion Out Loud language and another using traditional reproductive rights language. If the Abortion Out Loud messaging is more effective at generating actions, we can infer that for this universe of people, this message via email motivates them to take action. We can NOT infer that Abortion Out Loud is always a more effective message.

Each test should have a clear hypothesis and outcome measurement. Sometimes this means you must have only small differences between two versions to test a hypothesis. Again, in our example above - if in addition to different messages, we include different action asks or different subject lines, we can not isolate the factor causing the effect. In this scenario, it is impossible to know whether the abortion out loud messaging, the action or the subject line caused a change in action rates.

Variations within a test should also be distinct from each other - tactics that are too similar are unlikely to yield any results. If we wanted to compare the "Take Action" button on two identical landing pages, it would be vital that the buttons look quite different: for example, instead of comparing two shades of blue buttons, we should compare a bright red to a bright blue button.

Calendaring: Creating a calendar is vital to delivering testing content on time as well as to build a program that learns from past tests. Ideally, a calendar of tests for up to a few weeks will be regularly updated and maintained. The calendar should identify the test type (subject line, landing page), the variations (message A vs control) and the dates. If your messaging requires the creation of other online aspects, such as a donation page, or legal approval of messaging, include that in your calendar as well.

Interpreting results: Test results can surprise us, but we should have an outcome of importance in mind before a test is run. This helps us stay focused on the thing we are most interested in improving upon. A classic example is email open rates vs click through rates. Open rates are important because if no one is reading your emails, no one will take the call to action. But what we typically care about more is the action rate - how many people acted and therefore helped us achieve our programmatic goal? Another common measurement is unsubscribe rates - it is important to decide your threshold for list loss versus action - your subscribers that stay may be more productive and therefore worth a list size decrease.

For A/B tests, results are rarely generalizable. A tactic might work at a particular time in a particular situation, but that doesn't mean it will always work in every situation. You should repeat tests many, many times before considering the results a "best practice." While it is great to learn from other

tests, if another organization discovers something through testing, you should not assume it will work for you, since each audience is unique. You should test the tactic as well.

However, you can feel reasonably confident about drawing generalized conclusions if:

- You have tested something a few times and always gotten the same result.
- The results were very strong in one direction: results have a magnitude of effect size and are statistically significant.
- You have no reason to suspect that the audience/circumstances were special.
- You think it is very unlikely that the result is due to novelty effects.

Retesting results: For some tactics, we may decide to test them again either with the same parameters or a slight change. Figuring out a good testing schedule is more art than science. In general, you may want to re-test more frequently if:

- The original test results weren't very robust.
- You suspect novelty was a big factor.
- You think there was something unusual about the time period of the original test. Examples: Web traffic was atypical in some way; Specific subgroups were unusually activated

Often you'll just move on to newer and better iterations and won't really re-test old controls. This is particularly true with web pages. It is usually more efficient to aim for continuous improvement instead of re-testing.

Sample A/B Test Design

Below we describe a common A/B test - comparing subject lines and measuring action taking rates. In this case, we'll imagine a scenario where we test 2 subject lines and a control group, in which we ask people to sign an online petition.

Experimental Universe & Conditions

- *Subject line A: Targets receive one version of a subject line; body of the email and ask are unchanged*
- *Subject line B: Targets receive another version of a subject line; body of the email and ask are unchanged*
- *Control: Targets do not receive email*

The universe for this test should be determined by the larger goals of the program. For example, if we are primarily interested in young people in a district, the test should also focus on young people.

However, if your primary program is focused in District A, you can potentially run your test in District B if you are optimistic that the message may be effective in both. An example of a message that could work in District A & B would be a petition to "Stop Trump" and both District A & B have similar partisan make ups. A less ideal message would be one focused at a particular legislator in District A that does not also represent District B - the legislator might have a particular reputation that can impact your results.

To determine your universe size, we recommend using Optimizely's A/B Test calculator. You must make three assumptions for your test.

- **Baseline conversion rate - this is the percentage of people you think will complete the outcome you are measuring in your control group.**
    - In this case, the baseline conversion rate is the percentage of people that we believe will sign the petition without an email (control group). This is likely to be quite low, if they only way we are advertising the petition is via email.
    - We will assume a 2% conversion rate.
- **Minimal detectable effect (MDE) - MDE is the minimal difference in outcome between 1 treatment and the control group that you expect.**
    - For our scenario, this is how much better you think think subject line A OR subject line B will perform over the control group. MDE's can be informed by past tests by your organization or others.
    - As you add conditions, your MDE does NOT change. You are comparing each condition to the control group.
    - We will optimistically assume a 15% MDE for this test. This means we want to detect a difference of 15% between the control and treatment. If the conversion rate is 2% in the control group, then we want to be able to measure if the treatment group has a difference of 2% * 15%, i.e., if the conversion rate in the treatment group is between 1.7% and 2.3% (2% +/- 2% * 15%, or 2% +/- 0.3%).
- Statistical significance - Optimizely defaults to 90%. Unless you have strong reasons to change this, we recommend leaving the default.

With our example of 2 treatment conditions plus a control group, with a 2% conversion rate for the control group and a 15% MDE between condition and control, Optimizely estimates a sample size of 36,000 per condition and control, for a total universe of 108,000 targets.

Note that for different A/B tests, the sample size varies widely based on what you are measuring and your assumptions. You may be able to run tests with smaller sample sizes, but the smaller the sample size, the less likely you are to be able to measure a difference in effect size.

Implementation Details

Many online tools now include an A/B testing tool. If so, the process to set up a test is fairly simple:

- Create two subject lines and one email - in this case the body of the email should be the same for both conditions.
- Build the petition page - the language and layout should be the same for both conditions.
- Make the list - pull a list of 108,000 emails for the test, focusing on a universe similar to the program universe.
- Set up the internal A/B testing with splits for control and treatment groups.

If your organization does not have an internal A/B testing tool, A/B tests can still be conducted. Optimizely provides A/B testing tools that integrate with many other CRMs. However you may need to advise with someone with some statistical knowledge to correctly prepare the target universes, randomize the data and analyze the results.

Outcome Measurement

After the emails are sent, you should evaluate the results of the A/B test. Typically new signups drop off after the first few days, so this can be a fast process.

Each condition where an email is received should be compared to the control group. As we laid out in the beginning, we want to measure the number of petition signers; ideally one subject line produces a larger effect. We may also want to evaluate unsubscribe rates and open rates as

additional performance measures; if either are particularly high or low, it may change our assessment of reusing this tactic.

If the effects for one condition are better (more petition signers and not too many unsubscribes), that is a good indication that you should use that subject line to email the rest of your target universe. It will result in the most petition signers.

The value of a testing plan is that once a process is in place, you can quickly test and iterate, and A/B testing can be inserted into many digital outreaches. Planning a program around testing should help increase your organization's efficacy and outreach efforts.