A/B Testing for Email Fundraising

Greg Tozzi, Max Ziff, and Taeil Goh

December 13, 2020

Executive Summary and Recommendations

Thank you for the opportunity to work with the Children's Science Center. Below you will find a summary of our findings and recommendations. Details of our analysis are included in the technical report that follows this summary.

Research Questions We sought to answer two questions posed by the Center. Both questions were related to the effect on opening and click-through rates of treatments in emails sent by the Center soliciting donations ahead of the planned Fall 2020 fundraising surge.

1. Is there a difference in email opening or click-through caused by the choice of using the Executive Director's name and title in the from line versus using the Board Chair's name and title? *While we did observe a difference in opening rates favoring emails from Jill McNabb, the difference was not significant.*
2. Is there a difference in email opening or click-through behavior caused by the choice of using one of two potential subject lines? *The subject line, "You can be a Catalyst for STEM Learning," caused significantly higher opening rates.*
3. Were combinations of the two treatments particularly effective? *We found no significant effect due to the combination of treatments.*

We did not observe enough clicks to draw conclusions about the effect of the subject and from line choices on click-through rate.

Review of the Methodology The Center provided our team with a list of approximately 12,000 potential subjects for this study. This list was compiled from the Center's Altru database but specifically excluded individuals who were previously selected by the Center to receive a solicitation by traditional mail. We randomly selected 1,980 subjects to receive an email soliciting a donation. These individuals were evenly split into four groups of 495 with each group being assigned a unique combination of from and subject lines. The 68% of the population that had mailing addresses on file with the Center also had metadata provided by Boodle.ai related to their predicted affinity for various causes. We used this metadata to check the balance of our random assignment

The Center provided a final quality assurance check of the individuals we assigned to the study and removed 23 records. The Center sent emails to the remaining individuals using Mailchimp on October 20, 2020. We extracted the results from Mailchimp on October 28, 2020.

Results Results for open rates are summarized below.

We computed estimates of the effects of each of the two treatments. These are presented below.

Treatment	Effect (confidence interval)
Subj: You can be a Catalyst for STEM Learning	[0.004, 0.113] - Significant
From: Nene Spivey, Board Chair	[-0.064, 0.04] - Not significant

We did not find a significant effect resulting from the combination of the two treatments.

Applying the results We caution against drawing broad conclusions based on this experiment, particularly with respect to questions that the study was not designed to answer.

We are confident that the method we employed in this case produced evidence that the difference in opening rates was caused by the difference in subject lines. You can reasonably expect that this result would generalize to the remainder of the candidate email recipients during the current year-end fund drive. We suggest caution when drawing broader conclusions, however for the following reasons:

1. Responsiveness to the two subject lines may have been affected by factors that we could not control for, including the current public health and political situations.
2. We drew our subjects from a filtered list of potential recipients. Adding individuals to the population of recipients may invalidate the study's results.

The difference in opening rates between emails sent from Nene Spivey and Jill McNabb is interesting, but it is not significant. With the sample size we used, our experiment would not have detected a difference in opening rates smaller than roughly 5.5 percentage points. This is a fairly large difference. If you are interested in pursuing the question of which principal drives higher opening rates, we suggest that you run another experiment with a larger sample size.

If you desire to run additional tests in subsequent emails, or if you would benefit from a discussion on setting up a testing program, we are happy to set a call.

Introduction and Context

The remainder of this document is technical detail that would be supplied to the customer as an appendix.

The Children's Science Center The Children's Science Center ("the Center") is a non-profit organization headquartered in Fairfax, Virginia. Founded in TODO, the Center's principal objective is to build the Northern Virginia region's first interactive science center at a greenfield location on donated land in Reston, Virginia. The capital campaign to raise fund to build this facility is the Center's primary line of effort. A recent partnership with the Commonwealth of Virginia's

Recognizing the long-term nature of the capital campaign and the need to build interest and advocacy, the Center launched a series of intermediate efforts beginning in `TODO`. The Center's first outreach effort was the Museum without Walls, a program that delivered science-based programs to schools around the region. In `TODO`, the Center launched the Lab, a test facility located in a shopping mall sited 12 miles outside of the Washington, DC Beltway. The Lab hosted approximately `TODO` guests per year until it was temporarily shut down in response to the COVID-19 pandemic.

The Center funds operating expenses with Lab admissions, grants, and philanthropic giving. The Center's Development Department is responsible for developing and managing the latter two revenue streams.

The Fundraising Surge The Center conducts a fundraising surge toward the end of the calendar year, consistent with the charitable giving cycle in the United States.

The Center expected to engage the majority of its stakeholders by email. The Center viewed email marketing as repeated iterations of a three step process. The Development Director expected that several rounds of engagement would be necessary to achieve a conversion in the form of a charitable gift. Within each individual round, the path to conversion proceeded from receipt to opening to click-through using an embedded link or button.

The Center introduced two major changes to its process ahead of the 2020 fundraising surge. In previous year-end campaigns, the Center had used Constant Contact (www.constantcontact.com) to send branded emails to stakeholders. In the 2020 campaign, the Center chose to switch to Mailchimp's (www.mailchimp.com) free tier. The Center also engaged a consulting firm that applies machine learning to the nonprofit space to predict top prospects from among the Center's contact list and to predict affinities to aspects of the

Center's mission. The consulting firm's model requires a physical address as an entering argument, and it was, therefore, only applied to that subset of the Center's contact list for which physical addresses were entered. Based on the consulting firm's output, the Center chose to engage a subset of their stakeholder list with physical mail and to wall this cohort off from subsequent email engagement.

Research Questions The Center's development staff was most strongly interested in conducting experiments to maximize the opening rates of the Center's fundraising emails. The Development Director and her staff were keenly aware that email opening was the critical first step in achieving conversion. The follow-on actions—click through and conversion—were areas of secondary interest for this study. With this in mind, we sought to answer two specific questions posed by the Center.

1. Is there a difference in email opening or click-through caused by using the Executive Director's name and title in the from line of the solicitation email versus using the Board Chair's name and title?
2. Is there a difference in email opening or click-through behavior caused by the choice of using one of two potential subject lines? The two subject lines considered were, "You can be a Catalyst for STEM Learning," and , "Invest in the Power of STEM Learning."

Research Proposal

Research Hypotheses We claim expertise neither in non-profit email fundraising nor in the preferences of the Center's stakeholders. The Center's development team believed that there would be a significant difference in the opening rates associated with the two email from lines considered, though the staff did not have a sense of which would result in significantly higher opening rates. Underpinning the staff's expectation that we would find a significant difference was their belief that the Center's stakeholders had a meaningful sense of the Center's leadership and governance structure and would react differently when presented with emails from either the Executive Director or the Board Chair.

Both of the subject lines considered were written by the Center's development staff. The staff believed that subject lines could result in significantly different opening rates but did not have a going-in belief of which subject line would cause a stronger response.

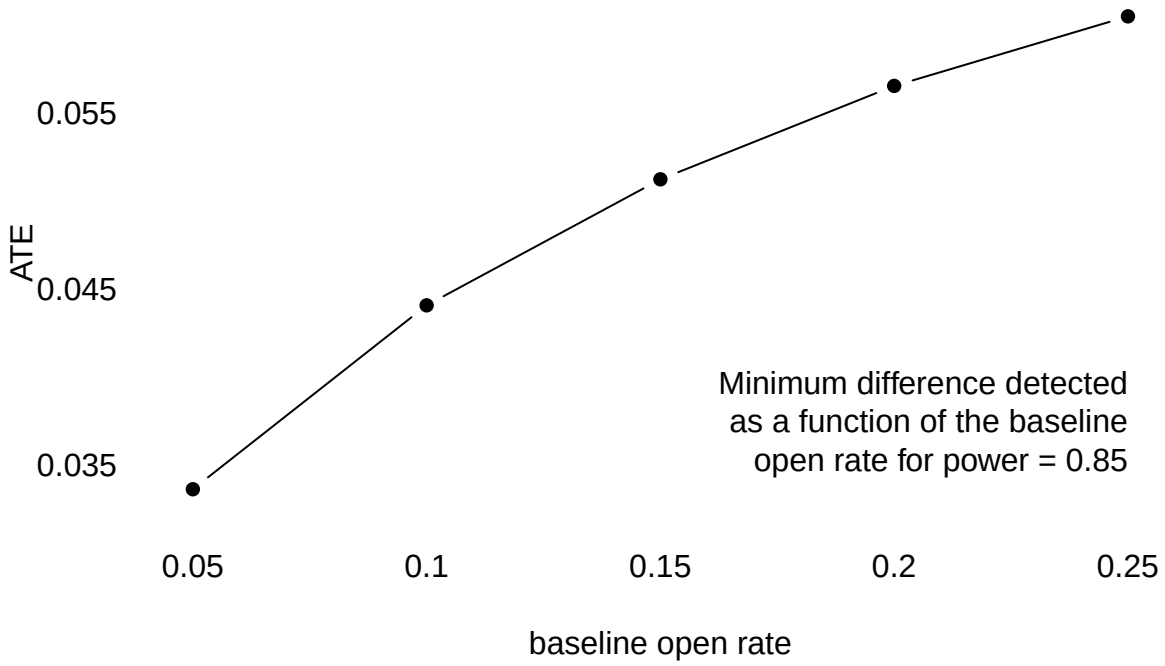
Experimental Design and Treatment in Details We addressed the research questions by construction a 2x2 factorial design using the difference in means estimator with the subjects assigned to one of four balanced groups:

- Group 1 - From Executive Director; You can be a Catalyst for STEM Learning - RX_1,1O
- Group 2 - From Board Chair; You can be a Catalyst for STEM Learning - RX_2,1, O
- Group 3 - From Executive Director; Invest in the Power of STEM Learning - RX_1,2O
- Group 4 - From Board Chair; Invest in the Power of STEM Learning - RX_2,2O

Our experiment follows the pattern While we are principally interested in the discrete effect of each treatment, this design allows us to explore heterogeneous treatment effects. Each group would receive a tailored email containing its assigned treatments sent through Mailchimp. We tracked opening and click through using Mailchimp out-of-the-box analytics. More detail about the implementation of the study is provided in the Outcome Measures section below.

Statistical Power The Center's experience across all of its email-delivered messaging suggested that we should expect opening rates on the order of 10%. To understand what deviations from this expectation would yield, we considered low, medium, and high baseline opening rates between 5% and 25% in our power calculations.

Powers between 0.8 and 0.9 are standard in clinical trials. We chose 0.85 as a target power and computed minimum detected differences for our range of baseline rates. Entering into the experiment, we believed that it would be unlikely that either treatment would produce differences in means as large as those that the power calculations suggested that we would need to report a significant finding.



Enrollment Process & Criteria for Subjects The Center's donor management database contains over 41,000 individual entries, roughly half which include physical addresses. Entries that include a physical address have metadata generated by a third party, Boodle.ai, a Data Science start-up that the Center engaged. These metadata were designed to predict affinities for causes central to the Center's mission, such as children's causes, educational causes, cultural causes, and scientific education. The Center provided us a file listing 12,004 individuals that Center intended to target by email during its Fall 2020 fundraising surge. These individuals represent all of the database entries that include email addresses less specifically excluded individuals (board members, minor children, etc.) and less a special cohort that the Center intended to target through a physical mail campaign based on Boodle's predicted capacity and willingness to donate.

The Center provided three data files that we used to develop our randomization. These files are located in the `data` directory in the project repository (www.github.com/gregtozzi/w241_project).

1. `anonymized_altru.csv` contains donor data for the Center's entire database of over 41,000 individuals. Entries are keyed to a unique donor identification number.
2. `anonymized_mail.csv` contains entries for a subset of the complete database for which the Center has physical addresses on file. This file includes third-party generated predictions of affinities for germane causes. Entries are also keyed to a unique donor identification number.
3. `anonymized_index.csv` contains the unique donor identification numbers for the approximately 12,004 individuals that the Center intended to target by email during the Fall 2020 fundraising surge.

We agreed to target a random sample of 1,980 individuals for this experiment. The intent was to remain within the bounds of the Mailchimp free tier while leaving a small margin for the Center's staff to send test emails to individuals outside of the experiment including to key members of the Center's staff and our group. The Center delivered one of four solicitation emails via Mailchimp's web interface.

To select subjects, we first filtered the extract of the Center's complete donor database on the list of individuals the Center intended to target by email. In the process, we removed three duplicated entries.

We had hoped to incorporate donor history into our randomization scheme and balance checks, but the data provided by the Center was extremely sparse and contained obvious errors. For every entry in the total gift amount column was either zero or `NA` and 21 entries in the first gift amount column were nonzero. The Center's staff was well aware of the data quality issues in their donor database. With zero covariates in the provided data, we turned to the output of the third party study contained in `anonymized_mail.csv`. We joined the data in that set to our existing table.

As explained above, outputs of the third party study only existed for individuals with mailing addresses on file. Only 8266 of the 12004 individuals in the population had data from the third party study as a result. We discuss the implications to covariate balance checks in the following section.

We conducted our randomization in two steps. First, we identified 1,980 individuals who would receive an email. Then, we randomly divided those individuals into four treatment groups of 495 individuals each.

lookup_id treatment 1: 1000 2 2: 1001 1 3: 1010 1 4: 1012 2 5: 1016 4 6: 1026 2

Validation of Randomization Procedure Data quality issues limited the number of covariates available to perform balance checks. In the end, we had membership data for our entire sample. We also had predicted affinities provided by Boodle.ai for about two-thirds of the sample. These affinities were for elements of the Center's mission: children's causes, cultural causes, educational causes, and science causes. Boodle's methodology requires a mailing address, so the remaining third of the sample did not have mailing addresses on file. Not having an address on file is an indicator of engagement with the Center, so we created an indicator variable capturing this covariate There were, of course, no instances in which an individual without an mailing address on file also had data in any of the Boodle-generated covariates. Including the no mailing address covariate in the balance check required that we run two separate regressions for each treatment variable—one which captured the Boodle.ai-generated affinities and membership data, and the other which captured the lack of a mailing address and membership data.

The regressions show no significant relationship between the covariates we tested and the treatment variables. We are confident that our randomization was successful.

Outcome Measures Implementing the experiment involved finding the proper balance between Mailchimp technical capabilities, the Center's expertise and availability, and the need to respect the individuals' privacy on the Center's mailing lists. Because the Center was new to Mailchimp, it was important to keep the implementation simple and inexpensive. This ruled out the use of Mailchimp's in-built experimentation features, both from the point of view of expense (experimentation features require paid subscriptions) and transparency: Mailchimp experimentation concealed their own randomization, which, while probably sound, can not be externally audited because it is proprietary.

The four treatment groups were modeled in Mailchimp by four different "campaigns", Mailchimp's basic data object for representing at least one email sent. Each campaign used a separate email template, with the appropriate variations hard-coded. A toy Mailchimp list and template we==re used to prove out the Mailchimp implementation. Also, in the live implementation, dummy addresses were added to each group that targeted the researchers, so we were reasonably confident that each group correctly received the email according to the experimental design.

Outcomes were gathered using a custom python script that in turn used Mailchimp RESTful API. During development, the script was run against the separate toy implementation to minimize the need to connect to the Center's live data, thus minimizing the risk of accidental leakage of private data.

Mailchimp's API provides a complete history of events for each send in each campaign. In particular, it distinguishes between these event types: open, click, and bounce. Mailchimp records multiple instances of each event type: a recipient may open or click on a link in an email multiple times, and each event is recorded. For our purposes, we gathered only the date-time of each event type's first occurrence, if any. As one would expect, click events were always preceded by an open event, and bounce events precluded any other events.

The Mailchimp API reports events keyed by email address. Thus it was necessary to run this script carefully, only on a trusted machine, and to then join this raw, sensitive data back to the full data set, with its

Table 2: Covariate Balance

	Dependent variable:			
	Subj: Catalyst		From: ED	
	(1)	(2)	(3)	(4)
Affinity - Children 35	−0.172		0.104	
	(0.276)		(0.278)	
Affinity - Children 90	−0.129		0.090	
	(0.286)		(0.290)	
Affinity - Education 35	0.104		−0.029	
	(0.198)		(0.199)	
Affinity - Education 90	0.110		−0.077	
	(0.180)		(0.180)	
Affinity - Science 35	0.082		0.013	
	(0.224)		(0.222)	
Affinity - Science 90	0.044		−0.034	
	(0.200)		(0.196)	
Affinity - Culture 35	0.025		−0.033	
	(0.047)		(0.047)	
Affinity - Culture 42	−0.006		−0.023	
	(0.074)		(0.074)	
Affinity - Culture 50	0.012		0.006	
	(0.078)		(0.078)	
Affinity - Culture 90	0.517		0.430	
	(0.321)		(0.882)	
Member	0.037	0.031	−0.001	0.001
	(0.029)	(0.028)	(0.029)	(0.028)
No address		0.028		0.010
		(0.026)		(0.026)
Constant	0.469***	0.484***	0.482***	0.497***
	(0.021)	(0.017)	(0.021)	(0.017)
Observations	1,356	1,980	1,356	1,980
R^2	0.003	0.001	0.006	0.0001
Adjusted R^2	−0.005	−0.0001	−0.002	−0.001
Residual Std. Error	0.501 (df = 1344)	0.500 (df = 1977)	0.501 (df = 1344)	0.500 (df = 1977)
F Statistic	0.416 (df = 11; 1344)	0.884 (df = 2; 1977)	0.704 (df = 11; 1344)	0.075 (df = 2; 1977)
Note:				*p<0.1; **p<0.05; ***p<0.01

covariates.

Accounting for Non-Compliance We observed two avenues for non-compliance. In the first, twelve individuals did not receive an email because they were removed from the study after being assigned to a group. The Center determined that these individuals could not be included in the study because they were minor children or were related to board members. Given the challenge of maintaining a stakeholder database with inputs across departments, we were very pleased that this form of non-compliance affected well under 1% of our subjects.

The second avenue for non-compliance was outdated or ill-formed email addresses. Mailchimp reported that 23 of our emails were not delivered. Again, we were impressed by the overall cleanliness of the data the Center provided.

To account for non-compliance conservatively, we coded the 35 non-compliers as having not opened the email. However, we believe that an argument could be made for simply neglecting these individuals from the analysis as the group of 12 who were removed should never have been included in the population, and the 23 who bounced would not have received the email in any event.

Research Report

Experiment Results The experimental results are shown formally in the table below. We first considered each treatment individually in columns 1 and 2. We consider a complete model of opening rates in column 3. While we recorded very few clicks, we included a model of clickthrough rates in column 4 for completeness. We report robust standard errors here and elsewhere.

Column 3 considers both treatments and the possibility of heterogeneous treatment effects. In comparing this model with those presented in columns 1 and 2, we note the following:

- Considering both treatments and the interaction term increases both the estimate of the effect of the use of the *Catalyst* subject line and its standard error. The estimate of the effect in the single factor model was [0.007, 0.082]. In the saturated model, this shifts and expands to [0.004, 0.113].
- We estimate the effect of the use of the *Board Chair* subject line treatment to be [-0.064, 0.04] in the saturated model. While not significant, the direction of the effect was not expected by the Center's staff, and this treatment may be worth further study with an appropriately-powered experiment.
- We did not find a significant heterogeneous treatment effect.

The last column examines click-through rates. For all treatments, there were not enough clicks to draw statistically relevant conclusions.

Generalizability and Mediation Analysis Although we saw statistically significant results in the context of our experiment, we are cautious to ascribe specific mediation mechanisms that would allow these results to be generalized. We feel that the specific variations we observed should generalize to the larger population from which we took our random sample. That is, if the Center had sent email to the larger group of 11k using the preferred subject line, we would expect that the improved open rate would apply to the entire population, but it is not clear that even that effect could be expected for hypothetical future mailings. We note that this experiment was run at a very specific moment, right before the Presidential election of 2020. There may have been special effects related to that time: the climate may have made the audience particularly responsive to certain keywords, or even more or less inclined to open mail at all.

The Center's development personnel have expressed that the lesson they draw from this is that email with a personal from-line out-performs email with a role from-line (e.g. "Nene Spivy" vs. "Executive Director"). This is interesting because this was not explicitly tested, and it illustrates the difficulty of translating intentions and desires of decision-makers into Data Science research questions.

It is interesting to speculate on what would be an interesting experiment or series of experiments to test a general hypothesis such as "email with a personal from-line out-performs email with a role from-line". We

Table 3: Effect of different subjects and senders

	Dependent variable:			
	Open Rate			Click Ra
	(1)	(2)	(3)	(4)
Subject - Catalyst	0.044**		0.059**	−0.00
	(0.019)		(0.028)	(0.004
From - Board Chair		−0.026	−0.012	−0.00
		(0.019)	(0.026)	(0.004
Subject - Catalyst, from - Chair			−0.028	0.000
			(0.039)	(0.004
Subject - Invest, from - Director	0.220***	0.256***	0.226***	0.004
	(0.013)	(0.014)	(0.019)	(0.003
Observations	1,980	1,980	1,980	1,980
R^2	0.003	0.001	0.004	0.001
Adjusted R^2	0.002	0.0004	0.002	−0.00
Residual Std. Error	0.428 (df = 1978)	0.429 (df = 1978)	0.428 (df = 1976)	0.045 (df =
F Statistic	5.330** (df = 1; 1978)	1.860 (df = 1; 1978)	2.580* (df = 3; 1976)	0.667 (df = 3
Note:				*p<0.1; **p<0.05; ***

expect that findings such as that would have to be at least somewhat temporally specific: the base might change from year to year - hopefully, it would become larger - and so its preferences might also change. It seems that ideally, experiments would become an on-going part of a development program.