Pre-analysis plan: Stay Smart Online Alert Service

Policy problem, trial aims and research question

The Stay Smart Online alert service is an email subscription that sends alerts whenever a major cyber incident has occurred. Alerts tend to be about prominent and widespread scams, fake emails, data breaches, or malware.

The aim of the alerts are to provide timely advice to members of the public either to prevent them from falling victim to the incident, or helping them to rectify the problem if they have been a victim, by providing advice and next steps.

This trial aims to test the effect of different behavioural concepts to see which email design elements increase the likelihood that people will engage with the advice, and share it with their friends, family or colleagues.

Sample and randomisation

This trial is an individually randomised email experiment conducted online as part of the Stay Smart Online Alert Service. The sample for this trial is the entire subscriber base for the Stay Smart Online Service at the time of implementation (i.e. around a week prior to sending the emails, to allow for randomisation). This is roughly 54,000 individuals. Subscribers who sign-up to the service after the date of randomisation will not be included in the trial.

Subscribers will be randomised (using a script) into one of six groups, with a balanced allocation ratio, using complete randomisation. Another BETA staff member not directly involved in the project will verify the randomisation code.

Interventions

All subscribers to the SSO alert service will receive an email outlining prominent online security threats from 2019. The trial is a 2x3 factorial design, thus we will have two independent variables (IV), one with two levels, the other with three:

IV. A: Call to Share: Email will include either an explicit prompt to share the email for the benefit of others, or a set of typical share buttons.

IV. B: Visual Design: Individuals will receive one of three possible emails with differing visual design; a standard email, an email with enhanced icons or an email with different icons. Emails will differ only in the icons included.

The table below shows the notation used to refer to individual groups formed from our two IV's.

		Visual Design		
		Standard	**Icons 1**	**Icons 2**
Call to Share	**No Call to share**	A0B0	A0B1	A0B2
	Call to share	A1B0	A1B1	A1B2

Primary outcomes

All outcomes will be measured exactly one week after sending the emails.

There are two primary outcomes, both of which are binary and from which we will calculate sample proportions.

1. Sharing = 1 if an individual clicks on the link to share the email, 0 otherwise.
2. Engagement = 1 if an individual clicks on a hyperlink for more information, 0 otherwise.

Primary hypotheses

H1. A1 > A0 (sharing outcome only)

Emails that contain the Sharing call to action will result in significantly higher rates of sharing than emails which do not contain the sharing banner.

H2. B1 > B0 (engagement outcome only)

Emails that contain the first set of icons (timing related) will show significantly higher rates of engagement than emails that contain no icons.

H3. B2 > B0 (engagement outcome only)

Emails that contain the second set of icons (action related) will show significantly higher rates of engagement than emails that contain no icons.

H4. B2 ≠ B1 (engagement outcome only)

Emails that contain the first set of icons (timing related) will show significantly different rates of engagement than emails that contain the second set of icons (action related).

H1, H2 and H3 will be one-sided hypothesis tests. H4 will be a two-sided test as we have no theory in regards to directionality.

Method of analysis

We will use ordinary least squares (OLS) regression to estimate our three main effects. These estimates, confidence intervals and p-values will be derived from a model with the following specification:

$$y = \alpha + \tau_1 A + \tau_2 B + \tau_3 C + \epsilon$$

Where y is one of our two primary outcomes, α is the intercept, τ_1 is the main effect of including a call to share, τ_2 is the main effect of including the first set of icons and τ_3 is the main effect of including the second set of icons, and ε is an error term which picks up variance not explainable by treatment indicators.

We will also estimate a second model which will include interaction terms for $A \times B$ and $A \times C$. We do not expect interactions, and have not powered the trial to detect them. Any evidence of strong interaction effects will be incorporated into our interpretation.

Secondary/Exploratory Analysis

All treatment arms have new functionality to print and/or save the email as a PDF version. We will apply the specified hypothesis tests and models to a secondary outcome, which looks at what effects our treatments have upon the rate of printing or saving the emails.

Sample size and power: splitting into testing and hold-out data sets

We expect a sample of roughly 54,000, which gives approximately 9,000 individuals per group. However, we expect only 48% of these will open the email and receive treatment. We will measure email open rates and only include those who open the email in our analysis. There is no plausible mechanism by which treatment could influence open rates. After this exclusion we expect a final sample of around 25,900.

We will randomly divide the data into a ‘testing data set’ and a ‘hold-out data set’ at a 1:1 ratio. (Note: we will only do this if the open rates are at least 30%. Otherwise, we will proceed with analysis of the entire data set). Splitting allows us to replicate findings for pre-specified hypotheses and conduct exploratory data analysis on the ‘testing’ data and update our hypotheses before analysis the hold-out data.

To implement the data splitting and analysis, we will follow these steps:

- One analyst will split the sample randomly and provide one half of the data (the testing data) to the main analyst, while keeping the hold-out data in a separate, restricted location.

- The main analyst will use the testing data to (a) test hypotheses pre-specified in this plan, and (b) conduct exploratory analysis to develop further hypotheses.
- Any additional hypotheses will be added to an updated version of the pre-analysis plan before the hold-out data is released to the main analyst.
- The hold-out data will be used to (c) replicate any results from the testing data, and (d) seek to confirm any new hypotheses generated from the testing data.

We will use an alpha level of 0.01 for all hypothesis tests.

Engagement

In previous emails, the rate of hyperlink clicks (among those who opened emails) was around 8%. For main effects relating to engagement, we estimate that we have 95% power (in both testing and holdout samples) to detect an increase in clicks from 8% to 9.8% at an alpha of 0.01.

Forwarding

We do not have data on previous forwarding rates. Assuming a baseline forwarding rate of 2%, at an alpha of 0.01, we estimate that we will have 95% power (in both testing and holdout samples) to detect an increase in forwarding from 2% to 3%.

Interpretation

We will make use of p-values to aid the interpretation of our results. However, we will consider the p-value together with effect size, robustness checks and design limitations to assess the strength of a finding.

Including a call-to-share or altering the icons of an email are very low cost interventions, meaning that a small effect could be practically meaningful. The intervention is low risk, so there is little consequence in acting upon a false-positive result.

Trial Threats

There is a chance that individuals who are signed-up for the service with multiple emails will receive both control and treatment group emails. However, we do not expect this to be a significant number of individuals.

We are unable to measure the amount of time that participants spend on reading each email. Instead we are using hyperlink click throughs as a proxy for engagement. In addition, we will be unable to measure the number of emails forwarded through the email server, rather than our forwarding banner. It is possible that the platform of email delivery will allow us to use a proxy measure for this (by

tracking engagement from users who are not registered to the alert service and therefore likely had the email forwarded to them). We will delve into this further in exploratory research; but regardless, it is a limitation of our ability to infer effects.

Missing Data and Exclusions from Analysis

Participants who unsubscribe from the Alert service will be recorded as missing data and excluded from analysis.

Considering our treatments will not have any effect on the likelihood that an individual will open the email we send them, individuals who do not open the email and did not engage with the email in any way will be excluded from our analysis.

However, users may be able engage with links and content in the emails without technically ‘opening’ them (mostly through email preview functions, which are quite common). Individuals that engage in this fashion but did not technically ‘open’ the emails will *not* be excluded from analysis.

Pre-analysis plan commitments

No analysis of trial data has been undertaken prior to the completion of this pre-analysis plan. We will also be transparent about, and provide justification for, any deviations from this plan.

Pre-analysis plan update: Stay Smart Online Alert Service

9 April 2020

Background and purpose

We used a 'Test' data set to test several pre-specified hypotheses and to conduct further exploratory analysis – see the original pre-analysis plan (PAP) for details. This PAP Update pre specifies changes to our hypotheses – based on the results from the Test data – before we commence analysis on the 'Hold Out' data set.

Email open rates

Contrary to our expectations in the PAP, we observed material differences in open rates between emails with icons and those without. These differences also appear to be statistically significant (to the extent that such a claim is meaningful for a hypothesis generated after the results are known). We speculate that some people may see the email in preview mode before actually opening the email. Some of those who see the icons during the preview may be more inclined to go ahead and open the email. This leads to two changes to our original PAP for our analysis of the Hold Out data.

First, the original hypotheses H1-H4 will be tested on the full data set, not just the subset of the data with those who opened the email. That is, we will revert to conducting a pure Intent To Treat (ITT) analysis. Since the outcome variables, expressed as percentages, will be much smaller for the full data set than for the subset who opened the email, we may, for convenience, present our outcomes as a fraction of 1,000 emails sent (rather as a fraction of 100).

Second, we will add a new outcome variable (email open rates, a binary variable = 1 if the email is opened, =0 otherwise) and a new hypothesis:

H5: B1 + B2 (pooled) > B0 (email open rate outcome). Emails that contain either icons (timing or action related) will show significantly higher open rates than those that contain no icons (one sided test).

Factorial design and interaction effects

In the original PAP, we pre specified that we would conduct tests using both a short-form model (without interactions between the two factors in our factorial design) and a long-form model (including interactions), both using OLS regression. The results from the Test data show material differences in both effect sizes and p-values between the two models. Consequently, we will follow the recommendation of Muralidharan, Romero and Wuthrich (2020) and use the long-form model with interactions for our main analysis for all hypotheses. We may conduct several of the following robustness checks:

- Remove the possible interaction by dropping treatments from the analysis (eg test H2 through a comparison of A0B1 > A0B0)

- The short-form model without interactions

- Logistic regression of the models above

Results for existing hypotheses

In the Hold Out data, we will re-test all existing hypotheses using our preferred method above (ie, using the full data set including un-opened emails, and using the long-form model with interactions). Here are our updated expectations based on the results from the Test data.

H1 (call to share): The results from the Test data confirm H1. We expect to replicate this result in the Hold Out data.

H2 (timing icon): The Test data results appear to confirm H2 although this conclusion is based on weighing the evidence from several analyses. Nonetheless, we expect to confirm H2 when it is re-tested.

H3 (action icon): The Test data gave a null result for H3 (both a large p-value and a negligible effect size). Based on this, we now think it less likely that the re-test will give a confirmation of H3.

H4 (timing icon vs action icon): The Test data gave a null result for H4 and so we expect the re-test will probably also return a null result.

Extension of H2

If H2 is confirmed in the Hold Out data, we will conduct further analysis to see if we can determine what type of engagement increased as a result of the timing icon. Preliminary analysis of the Test data suggests that the increased engagement mainly related to clicking on the ‘print and save’ link so our hypothesis, contingent on confirmation of H2, is that the timing icon increases engagement on ‘print and save’ more than on other engagement options.

References

Muralidharan, Romero and Wuthrich (2020) ‘Factorial designs, model selection, and (incorrect) inference in randomized experiments’ Working Paper (5 February) https://econweb.ucsd.edu/~kamurali/papers/Working%20Papers/CrossCuts%20(Current%20WP).pdf