# SAS 2022 Robocall Experiment

Written by: Brennan Hall, Caroline Dennis, and Katrina Washington
9/29/2022

## 1. Experiment Design

Current economic surveys and the Economic Census use a variety of data collection methods, including Telephone Follow-Ups (TFU), emails, and mail follow-ups, all in an effort to reduce nonresponse. In preparing for the 2022 Economic Census, we have been researching how to incorporate new technology into our follow-up efforts effectively. When our phone centers were shut down due to COVID-19, automated Robocalls were used to contact respondents. With increasing budget challenges, ECON decided to further explore the use of Robocalls for nonresponse follow-up in lieu of traditional Telephone Follow-up (TFU). Traditional TFU is conducted by a staff of individuals in a call center. Wages and overhead, among other factors, can make this a costly operation. Because of the expense, for the Service Annual Survey (SAS), only approximately one third of nonresponding cases are selected to receive TFU. For the 2021 SAS, while employing the use of Robocalls on cases that were not selected for TFU, we tested varying the timing of these follow-up Robocalls when used in conjunction with other follow-up efforts[1]. Cases that were not selected for TFU were assigned to two panels: the first panel received a Follow-up Robocall approximately one week after the mailing of the second Mail Follow-up, and the second panel received a Follow-up Robocall approximately three weeks after the mailing of the second Mail Follow-up. The Follow-up Robocall referred respondents to the recent letter, which was expected to support the legitimacy of the automated call. Cases were assigned to the two panels using the following conditions:

- Cases were assigned to panels after TFU selection
- In addition to cases selected for TFU, the following cases were determined to be ineligible for this experiment: Referral cases, Odyssey cases, Full Service Account Manager cases, May births, and cases without phone numbers.
- The remaining eligible cases were assigned to panels using a systematic sample using certainty vs non-certainty and industry as control variables; all cases within the same mailgroup or alpha received the same panel/treatment.
- Panel assignment was denoted using the _LI variable.

Table 1. *Panel Assignment Counts for the 2021 SAS Robocall Experiment*

| **Panel/Treatment (_LI)** | **Follow-up Robocalls** | **Case Counts** |
|---|---|---|
| A | FU Robocalls 1 Week after second Mail FU | 11,091 |
| B | FU Robocalls 3 Weeks after second Mail FU | 11,075 |

The results of this experiment will seek to answer the following questions:

1. How does the timing of Follow-up Robocalls impact:
    a. Overall response?
    b. Timeliness of response (how many days past mailout to receive a response)?
2. Is the impact on overall response or timeliness of response statistically different for Robocalls vs TFU?

[1] See Appendix A.1 for a full schedule of all contact strategies.

## 2. Preliminary Results

Since both panels received the same collection strategies from Initial Mail through the second Mail Follow-up, this paper will discuss the initial results for the data available from 2nd Mail Follow-up (May 9, 2022) through July 22, 2022, for a total of 74 days. This includes response data from 22,166 of the 22,191 cases that were selected for this experiment on May 10, 2022. There were 25 cases from the original 22,191 assigned to a treatment with a check-in date on the day of TFU selection (May 9) that subsequently were not included in the analysis. For this research, a response is defined as a case with a valid check-in date and a response code equal to "Y". Cases with a valid check-in date but a response code not equal to "Y" were treated as non-respondents when analyzing response. In this report, we will be discussing results only based on response code to define the probability of response.

We will first discuss a probability estimate model for a company's response using the unweighted company data. Ideally, a median estimate would be preferred for robustness, but at least 50% of total possible respondents are needed to respond in order to calculate a median estimate. Since these are companies requiring additional follow-up, the overall response rate and per treatment response rates are expectedly low (below 20%) in this dataset so a median estimate for the number of days until response is not calculable. Instead, a restricted mean is provided. We obtain relative response time information from a Cox Proportional Hazards Ratio model to provide a directional conclusion of the effect of sending an earlier robocall.

The comparison to companies that received a Telephone Follow-up (TFU) is analyzed at the end of this section.

### a. Probability Model for Response

We use a Kaplan-Meier probability estimate commonly used in survival analysis for clinical or community trials where a company's response represents the company's "death" or "exit" from the survival study.

For the unweighted company responses, we see in Table 2 there is an estimated 19.7% response rate for Treatment A and a 18.8% response rate for Treatment B by the end of this initial data collection. Through the rest of this report, we will evaluate if this is a statistically significant difference due to the treatment effects. Meanwhile, in Table 3, the mean time to respond (restricted to the 74 days) is 63.9 days for Treatment A and 64.5 days for Treatment B.

Table 2. *Unweighted company response rates per treatment on 74-day timeframe*

| Treatment | Response Rate | 90% Lower Bound | 90% Upper Bound |
|---|---|---|---|
| A | 19.7 | 19.0 | 20.5 |
| B | 18.8 | 18.0 | 19.5 |
| All | 19.2 | 18.7 | 19.8 |

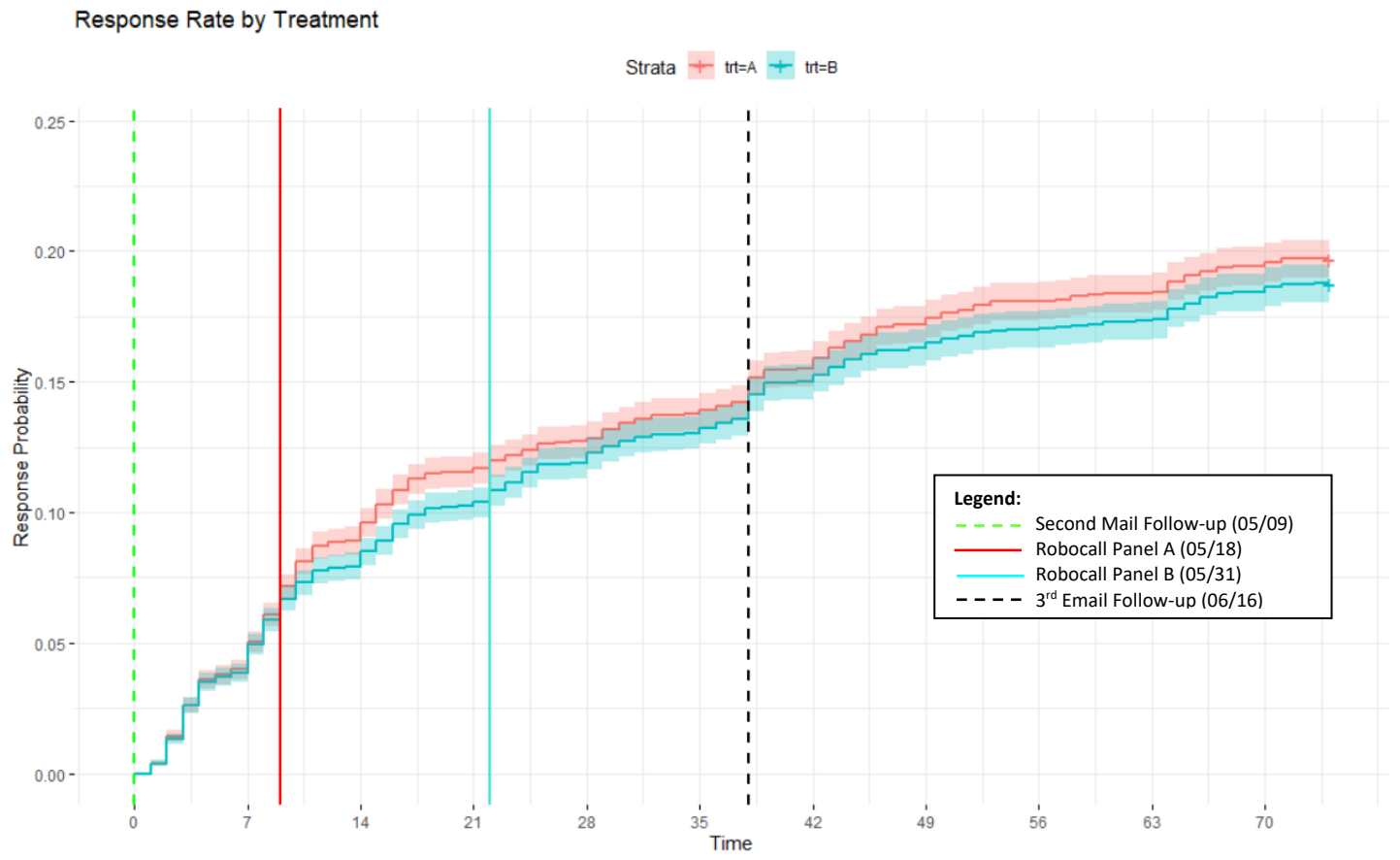
Table 3. *Unweighted company time to respond per treatment on 74-day timeframe*

| Treatment | N | Responses | Mean Response Time[2] | Standard Error |
|---|---|---|---|---|
| A | 11091 | 2187 | 63.9 | 0.211 |
| B | 11075 | 2079 | 64.5 | 0.205 |
| All | 22166 | 4266 | 64.2 | 0.147 |

[2] Denotes "Restricted Mean with upper limit of 74".

There is a noticeable divergence in the two response curves in Figure 1 after the Robocall for Panel A is made. The intent is to determine whether the response curve for Panel B makes a similarly significant return of convergence after the Robocall Panel B date which would indicate no meaningful effect on the timing of the robocalls.

Figure 1. *Response Rate by Treatment*




We investigate the difference in response distributions between treatments with a log-rank test which tests that the two response curves have equal distribution by comparing a function of their respective observed and expected event counts. The log-rank test results in Table 4 show a p-value of 0.07, indicating that the response curves for the treatment groups are statistically significantly different at a 90% confidence level.

Table 4. *Log-rank test for difference of response curve on 74-day timeframe*

| **Treatment** | **N** | **Observed (O)** | **Expected (E)** | **$(O-E)^2/E$** |
|---|---|---|---|---|
| A | 11091 | 2187 | 2127 | 1.68 |
| B | 11075 | 2079 | 2139 | 1.67 |
| $\chi^2$= 3.4 with 1 degree of freedom, p= 0.07 | | | | |

In addition to the difference in response rate, we also test the difference of timeliness of responses as reported in Table 3. By computing the 90% confidence interval of the two treatment's restricted means with the respective standard

errors, we get (63.85, 63.91) and (64.47, 64.52) for Treatments A and B, respectively. The confidence intervals do not overlap so we can conclude that the restricted mean of the response time for Panel A is smaller than Panel B.

An important recognition about this experiment is the potential impact of additional contact strategies biasing the result and confounding the efficacy of the Robocalls. As such, we consider shrinking the time frame of the experiment to only include days where treatments are directly related to the Robocalls, namely between the second Mail Follow-up and the day before the third Email Follow-up. This reduces the total time to respond for this experiment from 74 days to 38 days.

Analyzing this restricted timeframe, we find differing results than the 74-day timeframe analysis. Figure 2 illustrates the response curves for the two treatment groups for the abbreviated time period. The results in Table 5 indicate the difference in mean response time for Treatment A versus Treatment B can be concluded to have a statistically significant non-zero value following their respective (non-overlapping) 90% confidence intervals of (34.18, 34.42) and (34.48, 34.72). However, the log-rank test results in Table 6 indicates the two response curves are not significantly different at a 90% confidence level.

Figure 2. *Response Rate by Treatment on time restricted study on 74-day timeframe*

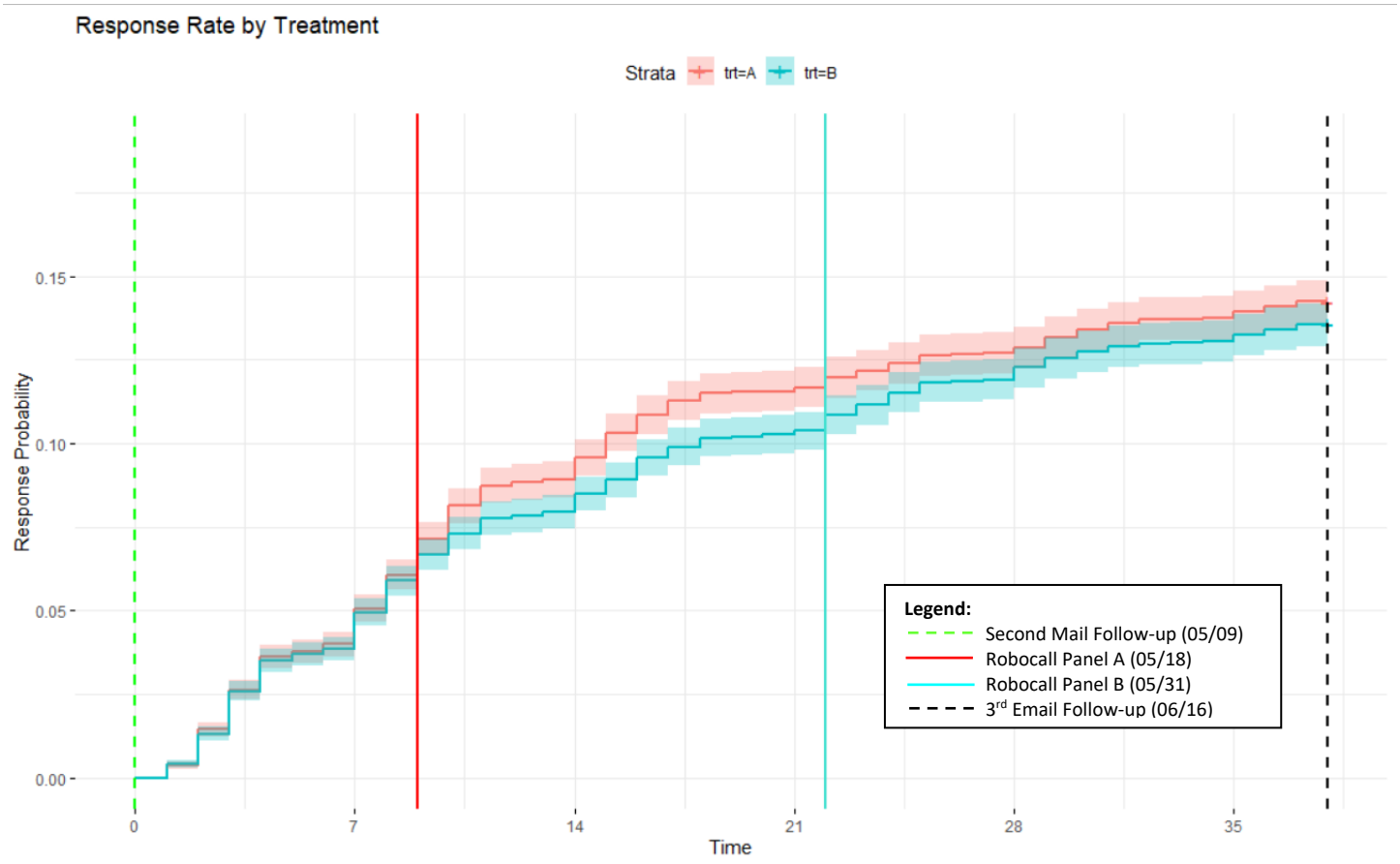


Table 5. *Unweighted time to respond per treatment on 38-day timeframe*

| Treatment | N | Responses | Mean Response Time[3] | Standard Error |
|---|---|---|---|---|
| A | 11091 | 1579 | 34.3 | 0.092 |
| B | 11075 | 1501 | 34.6 | 0.088 |
| All | 22166 | 3080 | 34.5 | 0.064 |

Table 6. *Log-rank test for difference of response curve on 38-day timeframe*

| Treatment | N | Observed (O) | Expected (E) | $(O-E)^2/E$ |
|---|---|---|---|---|
| A | 11091 | 1579 | 1537 | 1.17 |
| B | 11075 | 1501 | 1543 | 1.16 |
| $\chi^2$= 2.3 with 1 degree of freedom, p > 0.1 | | | | |

The log-rank test shows that there is no statistically significant difference in the response curves of the two treatments for the data that is restricted to the 38-day timeframe, which is inconsistent with the 74-day timeframe results. This could indicate that the follow-up contact strategies after June 16 may be confounding the results of the experiment.

Based on the results of the 38-day timeframe, the timing of the robocalls has a small but statistically significant impact on the mean response time and no statistically significant impact on the response rate. However, to thoroughly complete the experiment's analysis we continued with the analysis of the data from the 74-day timeframe since the difference in mean response time and the difference in response curves applicable to that timeframe are both statistically significantly different.

### b. Weighted Probability Model

We decided to consider a company's weight in our probability estimates to determine if there was an indication as to whether a company's size affected the response probability. By weighting the responders by the number of companies they represent in survey, and using the 74-day timeframe, we can see if the size of the company affects the treatment.

In relation to the hypotheses in question for this experiment, the two models yielded similar results and conclusions, so they are not reiterated here. Some figures and tables are presented in Appendix A.2.

### c. Cox Proportional Hazards Regression Model

To further confirm that treatment A has a significantly higher response rate, we model the hazard function in a Cox proportional hazards regression model. This model can be interpreted as the relative "hazard" (or "likelihood" in this research) for a company to respond with relation to different treatments. A noteworthy assumption for this model is that the hazard ratio between the treatment groups is assumed to be independent of time. We verify this assumption holds using the Schoenfeld test and provide those results in Appendix A.3.

For our Cox model, we choose to use Treatment B as the reference group, so our hazard ratio (HR) equal to 1.058 for Treatment A, as reported in Table 7, implies that around 1.058 times as many Treatment A companies are responding as Treatment B companies, at any given time. The proportional hazards model describes the effect of the earlier robocall as a small but statistically significantly non-zero positive effect on the response rate at a 90% confidence level, according to

[3] Denotes "Restricted Mean with upper limit of 74".

the associated p-value. When using weighted regression, the proportional hazards assumption is violated. This may indicate that the size of a company influences the results of the experiment.

Table 7. *Cox proportional hazards model results on 74-day timeframe*

| Covariate | Coefficient Estimate | Hazard Ratio | Coefficient Standard Error | P-value |
|---|---|---|---|---|
| Treatment A | 0.056 | 1.058 | 0.031 | 0.066 |

To ensure this low dimension model was not under-fitted and therefore biased, we built another Cox model with covariates that account for certainty status and industry. This likely overfitted model yields the same HR = 1.058 for the Treatment A effect, indicating the treatment effect is not due to any other observed underlying variables. The forest plot of this model's HRs is presented in Appendix A.4.

**d. Telephone Follow-up Analysis**

As part of our analysis, we also wanted to compare the cases that were selected to receive TFU to the cases receiving Robocalls. This allows us to analyze all cases that received some form of follow-up by phone. Treating the TFU cases as a control group for experimental purposes, we see the efficacy of TFUs versus Robocalls. The results for estimated response rate and mean response time are seen in Table 8 and Table 9 while the graph of estimated probability of response over time is given in Figure 3.

Table 8*. Unweighted company response rates per treatment on 74-day timeframe*

| Treatment | Response Rate | 90% Lower Bound | 90% Upper Bound |
|---|---|---|---|
| A | 19.7 | 19.0 | 20.5 |
| B | 18.8 | 18.0 | 19.5 |
| T | 22.6 | 22.0 | 23.2 |
| All | 20.7 | 20.3 | 21.1 |

Table 9. *Unweighted company time to respond per treatment on 74-day timeframe*

| Treatment | N | Responses | Mean Response Time[4] | Standard Error |
|---|---|---|---|---|
| A | 11091 | 2187 | 63.9 | 0.211 |
| B | 11075 | 2079 | 64.5 | 0.205 |
| T | 17310 | 3909 | 62.8 | 0.172 |
| All | 39476 | 8175 | 63.6 | 0.112 |

Similar to our initial analysis, we investigated the difference in response distributions between treatments with a log-rank test. The log-rank test results in Table 10 and Table 11 show a p-value less than 0.001 between the TFU group and Treatment Group A, and a p-value less than 0.001 between the TFU group and Treatment Group B, indicating that each pairwise comparison of response curves for Treatment A, Treatment B, and TFU cases are statistically significantly different at a 90% confidence level.

[4] Denotes "Restricted Mean with upper limit of 74".

Table 10. *Log-rank test for difference of response curve between Treatment A and TFU cases on 74-day timeframe*

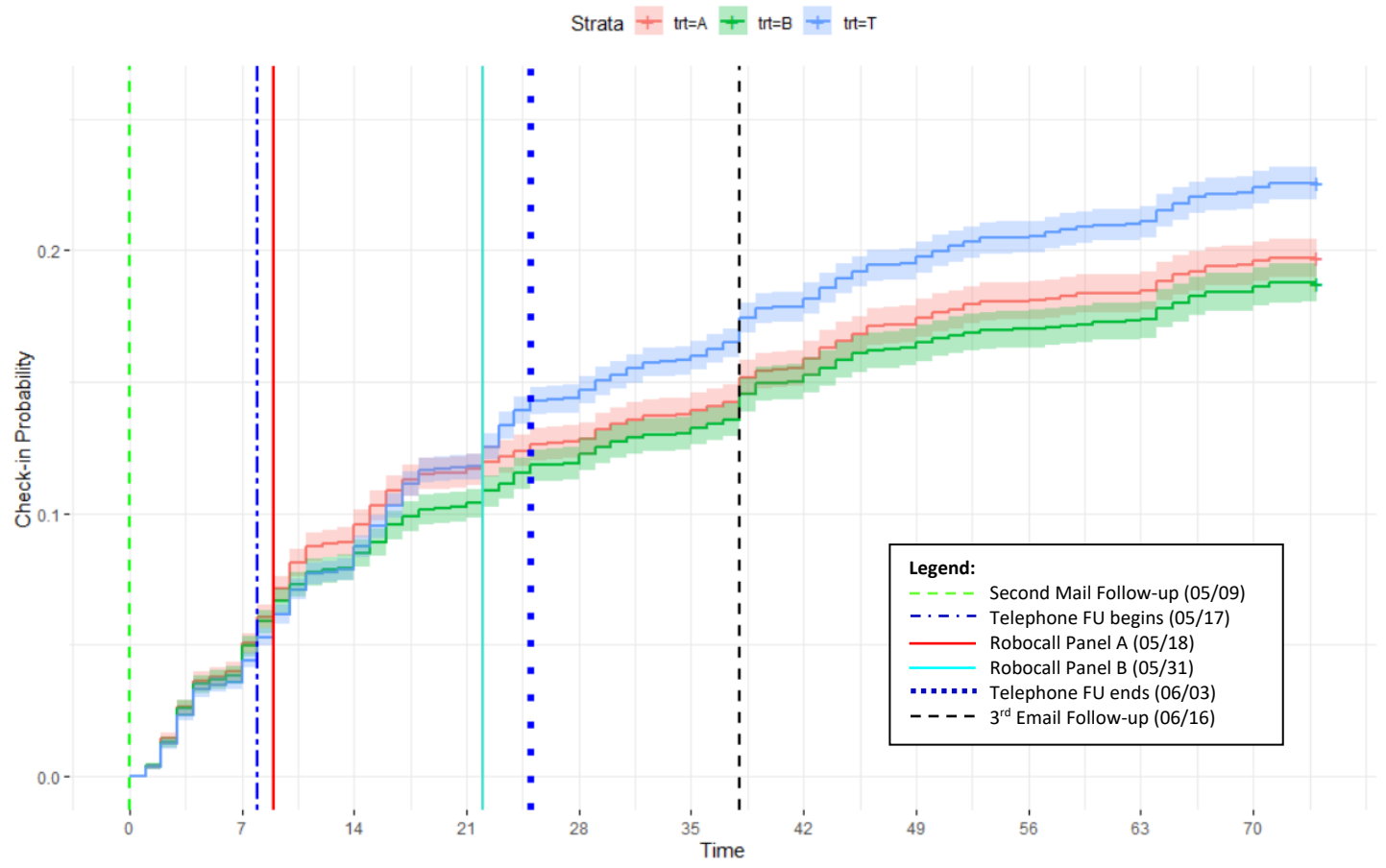
| **Treatment** | **N** | **Observed (O)** | **Expected (E)** | **$(O-E)^2/E$** |
|---|---|---|---|---|
| A | 11091 | 2187 | 2392 | 17.5 |
| T | 17310 | 3909 | 3704 | 11.3 |
| $\chi^2$=29.0 with 1 degree of freedom, p < 0.001 | | | | |

Table 11. *Log-rank test for difference of response curve between Treatment B and TFU cases on 74-day timeframe*

| **Treatment** | **N** | **Observed (O)** | **Expected (E)** | **$(O-E)^2/E$** |
|---|---|---|---|---|
| B | 11091 | 2079 | 2359 | 33.2 |
| T | 17310 | 3909 | 3629 | 21.6 |
| $\chi^2$=55.1 with 1 degree of freedom, p < 0.001 | | | | |

In addition to the difference in response rate, we also tested the difference of timeliness of responses as previously reported in Table 9. By computing the 90% confidence interval of each treatment's restricted mean response, we get (63.63, 64.17) and (64.24, 64.77) for Treatments A and B, respectively, and (62.58, 63.02) for TFU. The confidence intervals do not overlap so we can conclude that the mean time to respond for Treatment A, Treatment B, and TFU cases are all statistically significantly different at a 90% confidence level.

Figure 3. *Response Rate by Treatment with TFU cases*



We also analyze a Cox proportional hazards model with the TFU treatment as the reference group. However, upon conducting a diagnostic review of this model, we find the assumption that the hazard ratios are proportional over time no longer holds with the introduction of the TFU cases. This assumption is likely not met due to missing, unobserved covariates in the model.

## 3. Summary

The following is a brief review of the findings and conclusions from the experiment.

- When modeling responses from cases that received a Robocall treatment between Second Mail Follow-up (May 9) and Third Email Follow-up (June 16), the restricted mean time to respond for Treatment A and Treatment B cases is 34.3 days and 34.6 days, respectively. According to a confidence interval comparison, respondents receiving Robocalls just one week after the mail follow-up had a statistically significantly faster response time than the respondents who received Robocalls three weeks after the mail follow-up at a 90% confidence level.
- The estimated probability of response by June 16 for Treatment A and Treatment B cases is 14.2% and 13.6%, respectively. The difference in these probabilities is not a statistically significant difference at a 90% confidence level according to a log-rank test.

- When modeling responses from cases that received a Robocall treatment between Second Mail Follow-up (May 9) and the end of the initial data collection (July 22), the restricted mean time to respond for Treatment A and Treatment B cases is 63.9 days and 64.5 days, respectively. This difference is small but still statistically significant at a 90% confidence level.
- The estimated probability of response by July 22 for Treatment A and Treatment B cases is 19.7% and 18.8% percent, respectively. According to the log-rank test, these probabilities of response are statistically significantly different at a 90% confidence level.
- The treatment group is a statistically significant covariate at the 90% confidence level in the Cox proportional hazards regression model. Using the Treatment B as the reference group Treatment A has a hazard ratio of 1.058.

- Between Second Mail Follow-up (May 9) and Third Email Follow-up (June 16), the restricted mean time to respond for Treatment A, Treatment B, and TFU cases is 63.9 days, 64.5 days, and 62.8 days, respectively. All pairwise differences in the response times are statistically significant at a 90% confidence level.
- The estimated probability of response by July 22 for Treatment A, Treatment B, and TFU cases is 19.7%, 18.8%, and 22.6%. According to log-rank tests, all pairwise differences are statistically significant at a 90% confidence level.

# Appendix A.

## A.1. Contact Schedule

| **Collection Activities** | **Mail/Email Date** |
|---|---|
| Initial Mail | 02/10/2022 |
| Due Date Reminder | Letter: 03/10/2022 ***OR*** Email: 03/15/2022 |
| Due Date | 03/22/2022 |
| 1st Extension Reminder Email/ 1st Email Follow-up | 03/24/2022 |
| 1st Mail Follow-up (First Class) | 04/13/2022 |
| 2nd Extension Reminder Email/ 2nd Email Follow-up | 04/20/2022 |
| 2nd Mail Follow-up (First Class) | 05/09/2022 |
| **Follow-up Robocalls/ 1st Telephone Follow-up** | **Panel A: 05/18/2022 *OR* Panel B: 05/31/2022 *OR* 1st TFU: 5/17 - 6/3/2022** |
| 3rd Email Follow-up | 06/16/2022 |
| 4th Email Follow-up | 07/12/2022 |
| 2nd Follow-up Robocalls | 07/13/2022 (all cases) |
| 3rd Mail Follow-up (Priority Class) | 08/04/2022 |
| 5th Email Follow-up | 08/17/2022 |
| 6th Email Follow-up | 09/13/2022 |
| Closeout | 09/30/2022 |

## A.2. Additional Model Results

**Weighted Probability Model**

We decided to consider a company's weight in our probability estimates to determine if there was an indication as to whether a company's size affected the response probability. By weighting the responders by the number of companies they represent in survey, we can see if the size of the company affects the treatment. The tables below reflect similar results as the unweighted estimates analyzed in the body of the report.

Table A.1. *Weighted company response rates per treatment on 74-day timeframe*

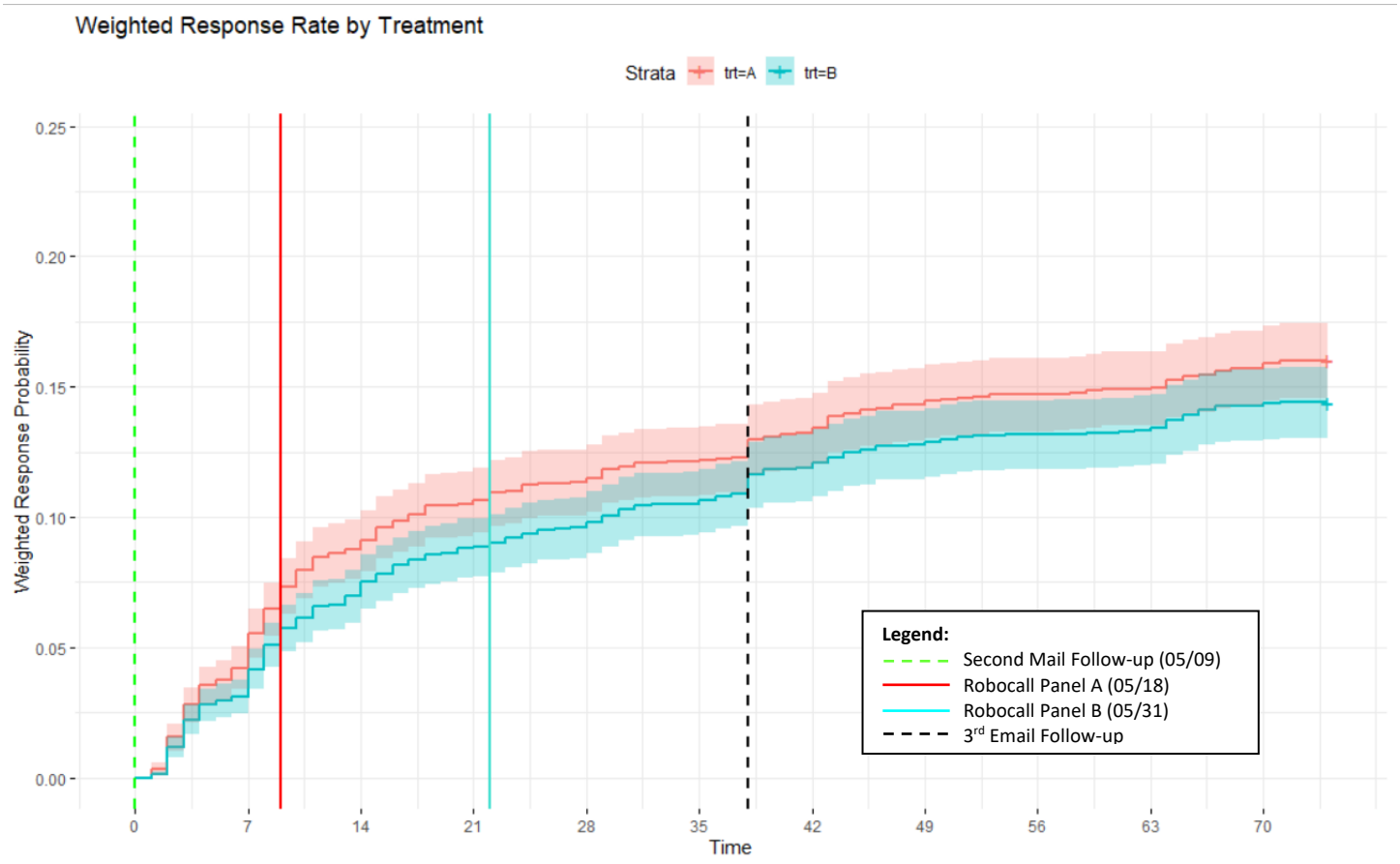
| **Treatment** | **Response Rate** | **90% Lower Bound** | **90% Upper Bound** |
|---|---|---|---|
| A | 16.0 | 14.5 | 17.4 |
| B | 14.4 | 13.0 | 15.8 |

Table A.2. *Weighted company time to respond per treatment on 74-day timeframe*

| Treatment | N | Responses | Mean Response Time[5] | Standard Error |
|---|---|---|---|---|
| A | 623571 | 99800 | 65.3 | 0.027 |
| B | 603394 | 86950 | 66.4 | 0.026 |

These results yield confidence intervals for the mean time to respond for Treatment A and Treatment B of (65.27, 65.33) and (66.37, 66.43), respectively.

Figure A.1. *Weighted Response Time by Treatment on 74-day timeframe*




A log-rank test for the difference of response curves yields:

$\chi^2$ = 661.1 with 1 degree of freedom, p < 0.001.

[5] Denotes "Restricted Mean with upper limit of 74".

## A.3. Model Diagnostics

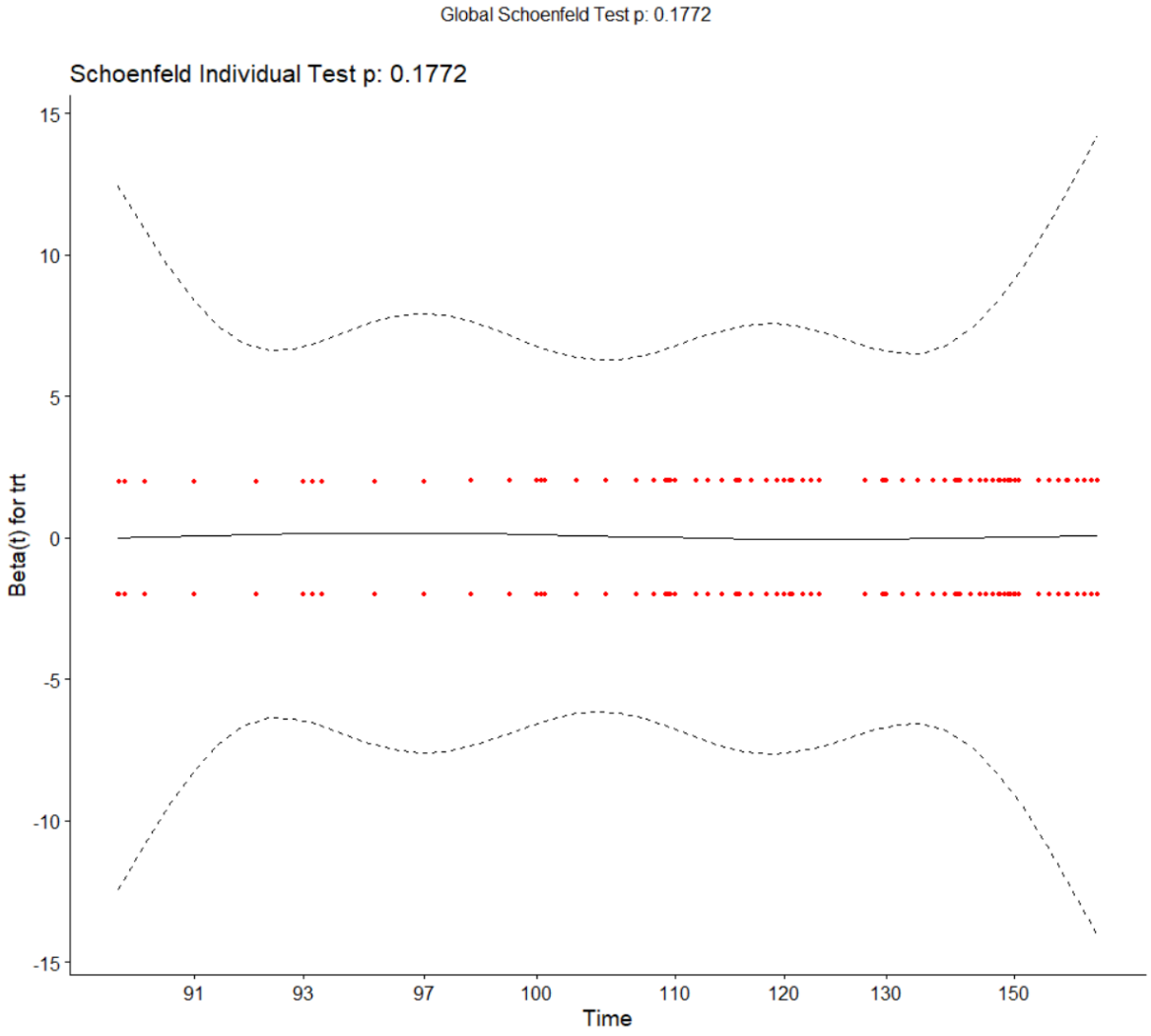
### Cox Proportional Hazard Ratio Test

The proportionality assumption is that the hazard rate of cases in both treatment groups is relatively constant in time. Rejecting the null hypothesis of the Schoenfeld test indicates assumption has been violated. In the unweighted model, the treatment covariate satisfies the assumption.

Table A.3. *Schoenfeld Proportional Hazard Test on 74-day timeframe*

| | $\chi^2$ | **Degrees of freedom** | **P-value** |
|---|---|---|---|
| Treatment A | 0.63 | 1 | 0.43 |

Figure A.2. *Schoenfeld plot of coefficient for Treatment, $\beta(t)$ vs. time*




With the introduction of the TFU cases, the proportionality assumption is violated because the null hypothesis fails to reject, as shown in Table A.4.

Table A.4. *Schoenfeld Proportional Hazard Test with TFU cases*

| | $\chi^2$ | **Degrees of freedom** | **P-value** |
|---|---|---|---|
| Treatment A | 18.11 | 1 | <0.001 |
| Treatment B | 6.51 | 1 | 0.011 |
| GLOBAL | 36.93 | 2 | <0.001 |

## A.4. Additional Cox Proportional Hazard Model Results

A model with treatment group, certainty status, and NAICS designation as covariates was modeled to determine any bias from confounding factors. The forest plot in Figure A.3 shows the distribution of the hazard ratios for each of the covariates.

In the figure, each row presents the estimate, confidence interval, and p-value for the estimated hazard ratios based on each of the categorical covariates. Of note is the first line corresponding to Treatment A which has a nearly equal hazard ratio of 1.06 as the hazard ratio presented in the main text of the report for the model with just Treatment as the covariate under the 74-day timeframe.

Figure A.3. *Forest plot of Cox Model with Treatment, Certainty, and NAICS covariates*