Experimental Design for Non-response Follow-up of an Electronic Questionnaire Survey

Joanne Leung,[1] Jeannine Claveau,[1] Claude Turmelle,[1] Karla Fox,[1] Wei Lin[2] and Milana Karaganis[1]

[1] Statistics Canada
100 Tunney's Pasture Driveway, Ottawa ON K1A 0T6, Canada
[2] University of Toronto
Ontario, Canada

Abstract

At Statistics Canada, business survey collection consists of many steps and modes of collection. Many business surveys continue to use paper questionnaires for data collection. Telephone follow-up is also conducted to resolve edit problems with mailed-back questionnaires, to remind respondents to return their questionnaires, and even to collect data from respondents who have not returned the questionnaires after a specific period of time. Currently, Statistics Canada is undertaking a general restructuring of its business statistics programs. One of its goals is to make electronic data collection the principal mode of collection for business surveys. Until now, follow-up methods used for electronic questionnaire collection were based on strategies used for paper questionnaire collection (fax and/or telephone reminders). Statistics Canada and other international organizations' experiences have shown that respondents of electronic questionnaires require different follow-up patterns. In order to establish a standard collection follow-up strategy for annual business surveys using an electronic questionnaire as a main collection mode, Statistics Canada built an experimental design to compare different non-response follow-up methods (i.e., combinations of telephone and e-mail reminders at different points throughout the collection period). An embedded balanced factorial design was used for this experiment. Analysis of paradata outputs—such as return rates, return time, and time spent on non-response follow-up—would permit us to compare the different methods. The experiment was aimed at providing evidence to establish the most efficient follow-up strategy, to assess the importance of having a first follow-up via a phone contact versus an e-mail reminder, and to know how far we can get if we use e-mail reminders only. This paper summarizes initial results of this experiment.

Keywords: Collection; Non-response; Follow-up; Paradata; Experimental design.

1. Introduction

At Statistics Canada, business survey collection consists of many steps and uses more than one collection mode. Many business surveys continue to use paper questionnaires for data collection. Recent developments in Internet technologies have greatly impacted survey data collection as the use of electronic questionnaires (EQ) for data collection has exploded over the last ten years. The EQ surveys could take on a variety of forms from simple e-mail surveys to sophisticated web survey systems. Statistics Canada is currently undertaking a general restructuring of its business statistics programs. One of the goals is to let electronic data collection become the principal mode of collection for business surveys.

The Unified Enterprise Survey (UES) program at Statistics Canada covers all the major industries in Canada (services, manufacturing, wholesale and retail) as well as some other smaller industries. It consists of close to 60 annual business surveys which employ a unified methodology in sampling, data collection, processing and estimation. Currently, the majority of UES business surveys continue to use paper questionnaires. Also, telephone follow-up is conducted to resolve edit problems with mailed-back questionnaires, to remind respondents to return

their questionnaires and to collect data from respondents who have not returned the questionnaires after a pre-specified period.

Until now, follow-up methods used for EQ surveys at Statistics Canada were the same as those used for paper-based surveys. These consist of a combination of fax and telephone follow-up attempts. International and Statistics Canada experiences showed that electronic respondents require different follow-up patterns. In order to establish a standard collection follow-up strategy for annual business surveys using EQ as a main collection mode, Statistics Canada built an experimental design to compare different non-response follow-up (NRFU) methods, combining telephone and e-mail reminders at different points throughout the collection period. Seven UES surveys in the 2011 collection cycle were used in this experiment. An embedded balanced factorial design was used in this experiment. Upon successful pre-contact of the sampled units, those who agreed to participate in electronic collection were randomly assigned to one of the four treatments, each with a different NRFU strategy implemented. By doing an embedded experiment, we aim at finding a follow-up strategy that produces the best response rates and is the most efficient. We would also like to know, in terms of return rates, how far we can get with sending e-mail reminders only, and to see if there is any importance of having the first follow-up attempt via telephone versus e-mail.

This paper summarizes the initial results of this experiment. Section 2 gives an overview of UES collection, including pre-contact, mail-out and follow-up of questionnaires, as well as a brief description of the paradata collected. Section 3 gives the methodology of the experimental design. Section 4 presents the results of the experiment, including descriptive statistics and results of analysis of variance tests. Section 5 summarizes conclusions and section 6 gives recommendations for future EQ surveys.

2. Overview of UES Collection

As mentioned, the UES program consists of close to 60 annual business surveys which are integrated in terms of content, collection and data processing. For most UES surveys, data collection takes approximately ten months. At the beginning of collection, a pre-contact is made in order to obtain up-to-date contact information of the enterprises in the sample. Then, questionnaires are mailed out to all sampled units on a specified date. The completed and mailed-back paper questionnaires are then imaged and captured. Each questionnaire is assigned a response code which allows us to classify them as 'respondent,' 'non-respondent,' 'out-of-scope' or 'still in progress.' NRFU begins one month after mail-out of the questionnaires. The intensity of the NRFU depends on a score function, which is based on weighted sampling revenue. Also, failed-edit follow-up (FEFU) is sometimes required to resolve edit problems found in the returned questionnaires.

2.1. Availability of EQ Surveys

In 2011, web-based electronic questionnaires were built and used for the collection of seven UES surveys. Six surveys were annual surveys of different service industries: Specialized Design, Accounting, Management Consulting, Engineering, Architecture and Software Development and Computer Services. EQ was also available for the Head Office Survey.

Before 2011, only one survey (Software Development and Computer Services) in the UES program was using electronic questionnaire as a collection mode.

2.2. Pre-contact

The first phase of the annual collection cycle for UES surveys consists of a telephone pre-contact made for new enterprises selected in the sample to confirm their contact information and activity codes based on the North American Industry Classification System (NAICS). Since the information for new enterprises provided by the Business Register may not always be correct, a pre-contact call is required to make any corrections to this information when necessary. The pre-contact is also important because it determines which enterprises should be excluded from the sample due to changes of ownership, structure or main business activity.

In 2011, for the six surveys using EQ as a collection mode for the first time, 100% telephone pre-contact was conducted not only to confirm existing information, but also to obtain e-mail addresses of the respondents.

Following the EQ push strategy at Statistics Canada, respondents were advised that collection would be done via the electronic questionnaires and were asked to supply e-mail addresses. The only respondents assigned to paper questionnaires were the ones that adamantly refused EQ or the ones that we did not reach during pre-contact.

For units in the Software Development and Computer Services Survey, only a subset of sampled units was pre-contacted. This included new units in the sample, paper respondents in the previous survey cycle that were still in sample in the current cycle, as well as non-respondents from the previous survey cycle that were still in sample in the current cycle. For the rest of the units in the sample, a heads-up e-mail was sent instead to advise that their responses would once again be collected through EQ.

2.3. Mail-out of Questionnaires

The next phase of the collection consisted of mailing out either paper questionnaires or e-mail invitations to the sampled units. For units that were assigned a paper collection mode, a paper questionnaire was mailed out on March 18, 2011. For enterprises that were assigned to an EQ collection mode, an e-mail invitation to complete the EQ was sent on March 28, 2011.

2.4. Follow-up

The final phase of collection for the UES surveys consists of receiving completed questionnaires, imaging paper questionnaires, uploading both imaged and electronic questionnaires into the central collection system (Blaise) for edit verifications, following-up with outstanding respondents (NRFU) and respondents whose questionnaire failed editing in Blaise (FEFU), finalizing of the cases in Blaise and outputting to the subject-matter division for further processing.

2.4.1. Non-response Follow-up (NRFU)

For the seven surveys using EQ in 2011 collection, follow-up on the units that did not submit completed questionnaires was subject to the design of experiment. For units with an EQ collection mode, four different non-response follow-up strategies (treatments) were designed. Units were randomly assigned different treatments. In each treatment, the units were followed-up according to a specified NRFU strategy if they had not responded at pre-specified points in time. The four different follow-up strategies for the treatments are described in section 3.2. Once the treatments were completed, interviewers conducted the final blitz on the outstanding units to further improve response rates.

Units responding via a paper questionnaire followed the standard NRFU strategy (telephone follow-up and/or fax reminders) as per the Score Function.

Regardless of the collection mode, all telephone NRFUs were allowed, a maximum of five attempts before a unit was considered as a final non-response unit.

2.4.2. Score Function

The Score Function is used to prioritize non-response follow-up (NRFU) for units using paper questionnaires. Typically, a NRFU strategy consists of a series of reminders / follow-up attempts for the units who have not returned their questionnaires by pre-specified time points in the collection period. Follow-up attempts could be done via fax or via telephone. All non-respondents of paper questionnaires are followed-up by fax. However, not all of them are followed-up by telephone. The Score Function measures the importance of the units based on their revenue contribution within each cell, which is a combination of NAICS, province and stratum. The revenue used is the revenue available in the Business Register when the sample of the survey is drawn. The top units whose combined contribution exceeds a predetermined threshold will receive telephone follow-up during collection, indicated by a follow-up flag of 1. The remaining units will not be followed-up by telephone, indicated by a follow-up flag of 0.

Before the mail-out of the questionnaires, initial scores are calculated. Throughout collection, the Score Function is adjusted dynamically twice a week according to the response codes of the units within each cell. When the cell has reached its threshold or when there are no more in-progress units, telephone follow-up stops for all units of the cell.

For details about the calculation of the Score Function, please refer to Evra and DeBlois (2007).

For 2011 collection of the surveys with an EQ collection mode, the Score Function was applied as per usual to all paper units. For units assigned to the EQ collection mode, initial scores were calculated as well. However, throughout the EQ NRFU, all EQ units were treated as if they were assigned a follow-up flag of 1, thus making them eligible for follow-up. Once the experiment was completed (as per the schedule described in section 3.2.), initial scores were reapplied to the outstanding EQ units to initiate the final collection step, namely, final blitz by interviewers to further improve obtained response rates.

2.4.3. Failed-edit Follow-up (FEFU)

Electronic questionnaires included some edits (e.g., consistency edits) to ask respondents to resolve obvious data quality issues while completing an online questionnaire. Once electronic questionnaires were submitted, they were uploaded into the central collection system, Blaise, for further collection edit verification. This was also done for paper questionnaires: once received, they were digitized and uploaded into Blaise. Interviewers reviewed all cases with edits triggered by Blaise. If interviewers could not resolve triggered edits based on available information, respondents were contacted over the phone. FEFU was conducted in the same manner and followed the same editing rules for all units, regardless of the collection mode. Hence, the embedded experiment focused on NRFU strategies only and FEFU was deemed out of scope of the study.

2.5. Paradata

Blaise provided the main source of paradata used to analyze the experiment. The Blaise System is used to record the history of all calls made to enterprises during pre-contact and calls made for non-response and failed-edit follow-up activities for the UES. Every time an operator accesses a collection unit, the system records a wealth of information. The data file generated by this system is called the Business Transaction History (BTH). Examples of paradata included in the BTH are collection unit identification, the date and the amount of time a case was open, the status of completion of an operational phase and relevant information about each call (e.g. the number of call attempts made, the result of the call, the appointment reason, etc.). With the availability of the BTH files, we were then able to compare the performance of different NRFU strategies.

3. Methodology of the Experimental Design

3.1. Embedded Experimental Design

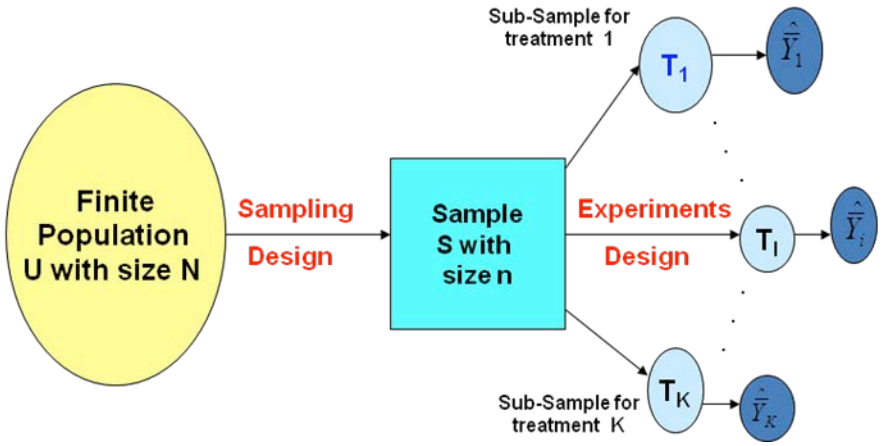
Researchers agree that it is important for survey methodologists to study how different survey methodologies and implementation strategies affect non-response, quality and efficiency (Van den Brakel and Renssen, 2005; Jackle et al. 2010; Groves and Lyberg, 2010). Randomized controlled trials are the most rigorous way of attempting to determine if a cause-effect relation exists between a specific intervention and an outcome. Other study designs can detect associations between an intervention and an outcome, but they cannot rule out the possibility that the association was caused by a third factor linked to both the intervention and the outcome. Although randomized trials are powerful tools, their use is often limited by ethical and practical concerns. In the case of the UES it would not be feasible, due to the cost, to set up an experiment separate from the survey to study different non-response follow-up strategies. However, recent research literature demonstrates that experiments embedded in ongoing sample surveys are particularly appropriate to investigate the effects of alternative survey methodologies on response behaviour or estimates of finite target population (Van den Brakel and Renssen, 2005; Van den Brakel and Van Berkel, 2002). Even if it turns out that the alternative approach has significant effects, we still could use the data collected from the current survey approach for publication (i.e. estimation) and use all the data for testing treatment differences (i.e. inference).

An experiment within a survey can be thought of as a variation of a two-phase survey design, as illustrated in Figure 1. As such, to test if the treatments are significantly different we need to use methods that take into account both the experiment phase and the sampling phase (Van den Brakel and Renssen, 1998).

For our experiment, as there were no previous estimates of effect size, we used an embedded balanced factorial design. Treatments were assigned in a balanced fashion randomly within survey strata. This randomization was done at the time of the survey design. This was done for operational reasons as randomization of non-responding units was not feasible in the current Blaise system framework. Randomization at this time point does allow for comparisons between treatments, but we need to study the effect of differential response from time of randomization on the treatment (Jackle et al., 2010).

However, a significant constraint was placed on our experiment by the fact that we used a live collection – we had to ensure that we had back-up plans to switch off the experiment if response rates were falling short of expected. In other words, we could not jeopardize the ultimate goal of producing estimates from this collection. Thus, we chose to close our experiment before the end of the collection period, to allow for the final blitz by interviewers to further improve obtained response rates.

Figure 1: Illustration of an experiment embedded in a sample survey



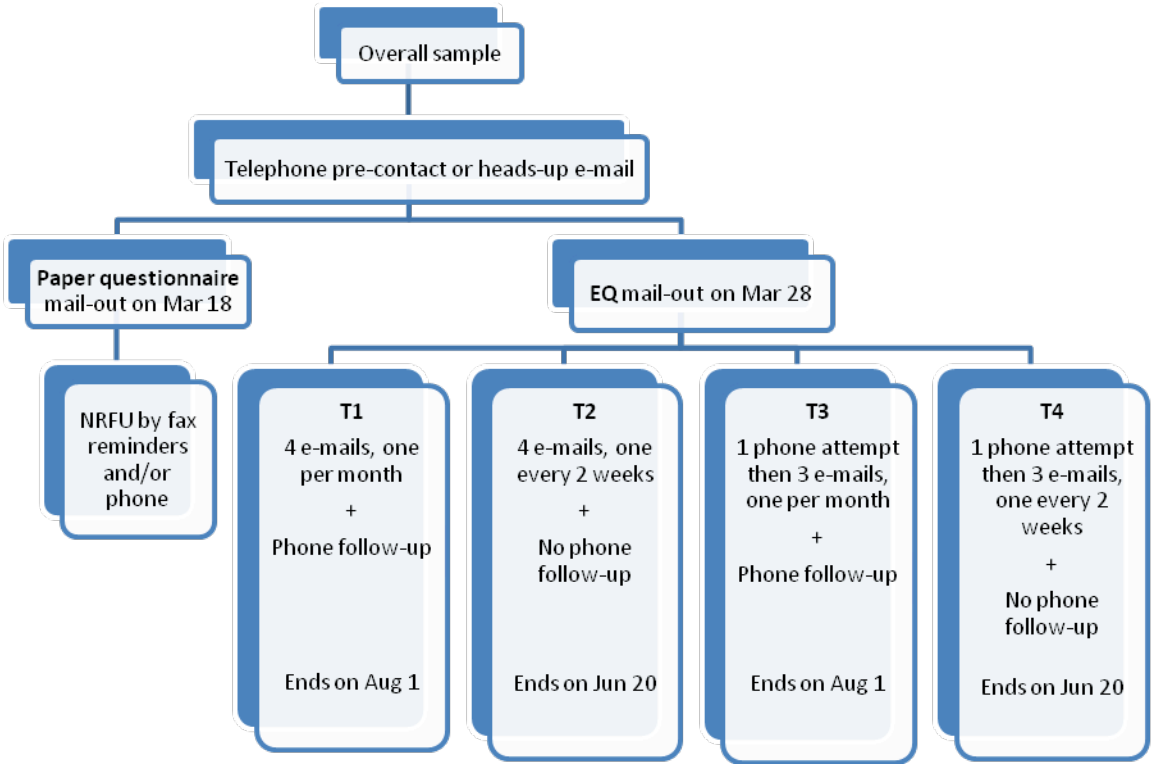
3.2. Treatments

The set up of the experiment is illustrated in Figure 2. Note that the split between units receiving a paper questionnaire and an electronic questionnaire was not performed by randomization. Instead, it is determined by the respondents. Therefore, we cannot make comparison on the results between paper and EQ collection. In fact, paper collection was composed of businesses that were not contacted during pre-contact and those who refused to use EQ. Thus, these businesses are more prone to be a non-response and are suspected to be qualitatively different from the units who went with the EQ collection mode.

Paper units followed the usual NRFU strategy applied to all UES surveys, with NRFU attempts performed over the phone as per score function, and fax reminders once a month. Paper questionnaires were mailed out on March 18, 2011.

Units with an EQ collection mode were randomly assigned to one of the four treatments, where each treatment contained approximately 25% of the EQ sample. The randomization was done within each combination of survey, strata group and type of questionnaire. The three strata groups are: (1) small take-some stratum, (2) large take-some stratum and (3) take-all and must-take stratum. The type of questionnaire was identified by its length, i.e., long form or short form. It has to be noted that stratification was applied to the six service surveys while the Head Office survey was a census. As a result, we ended up with a randomized block design for service industries (surveys and strata used as blocks) and a completely randomized design for the Head Office survey. Thus, results from the analysis of variance tests had to be analyzed separately for service surveys and the Head Office survey to account for differences in the design.

Figure 2: The set up of the experiment



On March 28, 2011, each unit in Treatments 1 to 4 (T1 to T4) received an e-mail invitation that contained a hyperlink and an access code for completing the survey online. Non-response follow-up started approximately one month into the collection, on April 26, 2011. On that date, all outstanding units were sent to NRFU by a specified approach indicated for T1 to T4. Each treatment had a different NRFU approach using a combination of e-mail reminders and phone attempts. NRFU stopped for a unit as soon as the questionnaire was returned to Statistics Canada.

For the EQ units, the focus was on how to set up formal reminders (like the ones done via fax for the paper collection). The four treatments were designed as follows. Treatment 1 ("standard") was designed to mirror the usual NRFU strategies for paper questionnaires. In other words, we allowed phone follow-up throughout the experiment and once a month (April 26, May 26, June 23 and July 25), we e-mailed a reminder to complete outstanding questionnaires (thus, substituting fax reminders with e-mail reminders). Treatment 1 was to be used as a baseline for comparison with other strategies. This treatment was scheduled to finish on August 1, to allow for the final blitz by interviewers.

Treatment 2 was designed to test how far we could get with e-mail reminders only, i.e., what kind of response rates we could achieve without any phone follow-ups. We did not allow telephone follow-up, but interviewers were allowed to answer phone calls from respondents and make firm appointments to collect data, if insisted by respondents. Because there was no phone follow-up at all, we decided to send e-mail reminders more often, once every two weeks (April 26, May 9, May 26 and June 7). Thus, this treatment was scheduled to close on June 20, to allow for the same number of e-mail reminders as the other treatments.

Treatment 3 was designed to measure an impact of doing the first follow-up attempt via phone rather than by e-mail. In this case, the first follow-up attempt was conducted over the phone and the rest of follow-up reminders were done via e-mail, once per month (April 26-29 (phone attempt), May 26 (e-mail), June 23 (e-mail) and July 25 (e-mail)). Phone follow-up calls were conducted, similar to Treatment 1. For Treatment 3, the experiment ended on August 1, again, similar to Treatment 1.

Treatment 4 was designed to combine both having first follow-up attempt done over the phone and then switching to e-mail reminders only with a compressed follow-up schedule of every second week (April 26-29 (phone attempt), May 9 (e-mail), May 26 (e-mail) and June 7 (e-mail)), similar to Treatment 2. As such, we did not allow telephone follow-up, but interviewers were allowed to answer phone calls from respondents and make firm appointments to collect data, if insisted by respondents, similar to Treatment 2. For Treatment 4, the experiment ended on June 20, again.,

At the end of the experiment (June 20 or August 1), all outstanding units were sent to telephone NRFU according to the initial scores calculated in the Score Function. Although for Treatments 2 and 4 the end of the experiment was on June 20, telephone NRFU did not begin until July 8 for these treatments. This is because initial scores could not be put back before July 7. Note also that these two treatments received another e-mail reminder on July 7. After August 7, the experiment was completed for all treatments and telephone NRFU blitz for all four treatments was in operation and the dynamic process of the Score Function was applied as normally.

Active collection continued until October 14, 2011 where all follow-up actions stopped. This marked the end of the collection cycle for 2011.

4. Results

4.1. Descriptive Statistics

There were a total of 9,324 units sampled for these seven surveys. Among these units, 6,457 units (approximately 70% of the overall sample) were assigned to EQ collection after pre-contact. These EQ units were then divided randomly into four treatments of approximately the same size. The four treatments had 1,615, 1,613, 1,615 and 1,614 units respectively.

Through follow-up of non-response, respondents could request a switch of a collection mode. Thus during collection, we observed 338 units switching from paper to EQ and 521 units switching from EQ to paper collection. Also, it was determined during collection that 1,098 were out of scope.

At the end of collection, the number of in scope units in the four treatments were 1,375, 1,353, 1,394 and 1,376 respectively.

4.1.1. Return Rate

Return rate is often used as a key measure of survey progress. The return rate indicates the percentage of questionnaires completed and returned. It is determined when the questionnaire is submitted to Statistics Canada or if the unit is declared a respondent via another collection mode (CATI, fax, …). Business surveys have highly skewed populations, meaning a relatively small number of units can account for a large portion of the economic activity. Therefore, return rates ought to be calculated both on a weighted and an unweighted basis. The unweighted return rate indicates the percentage of questionnaires received or completed, among all in scope units, whereas the weighted return rate is the percentage of the revenue contribution of the received or completed units, among the overall revenue contribution of the in scope units.

The progression of total unweighted return rate for each treatment from the EQ Mail-out (March 28) to the end of collection (October 14) is shown in Graph 1. The return rates were computed based on all in scope units at the end of collection. All the events that happened on the key dates are indicated by the dotted lines labelled from “E1” to “E8”.

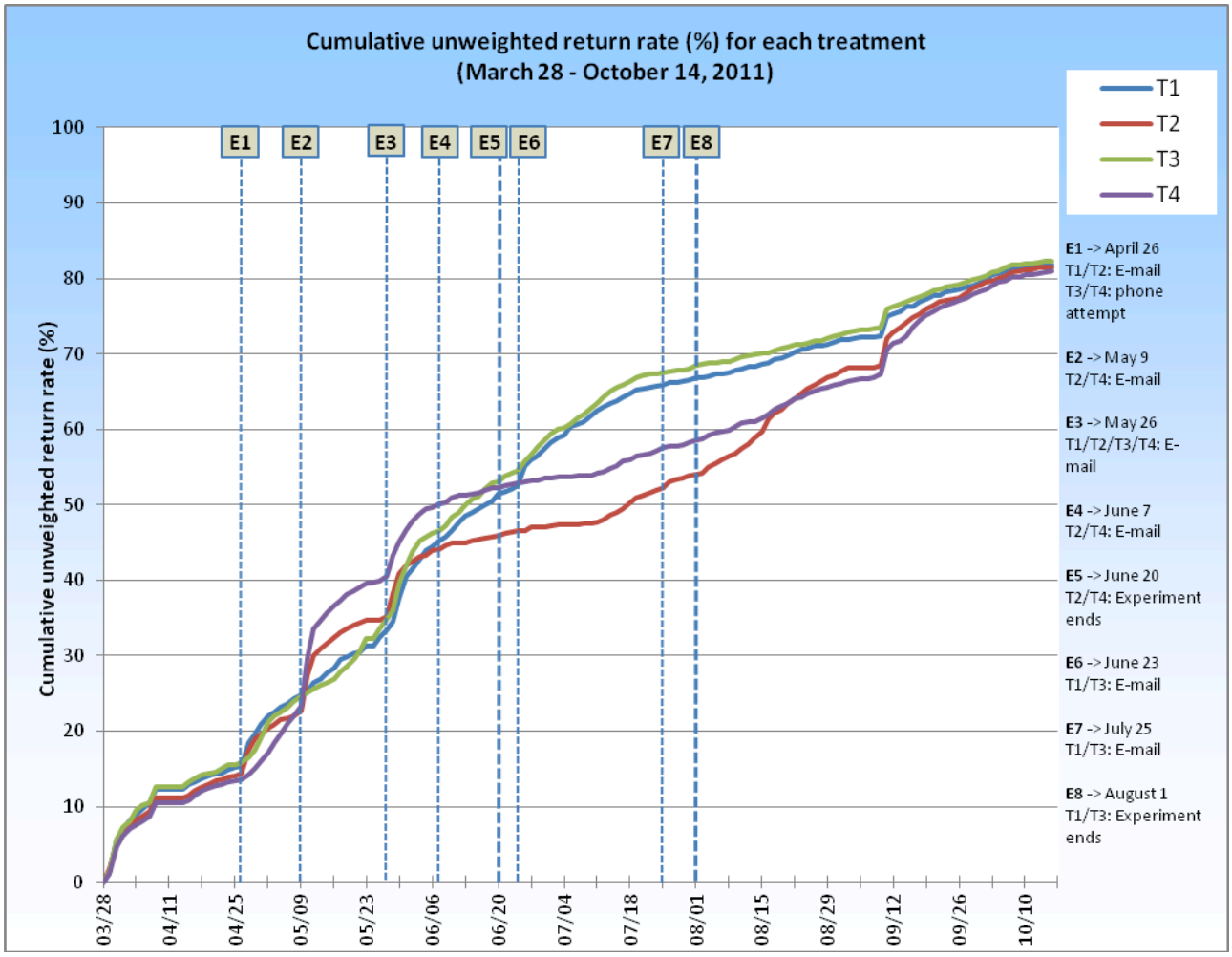
At the beginning of the non-response follow-up (April 26, E1), the unweighted return rate of the four treatments were similar (14% to 16%). On April 26, Treatments 1 and 2 received an e-mail reminder and Treatments 3 and 4 received a phone attempt in that week. Two weeks later (May 9, E2), all the treatments and paper collection had similar unweighted return rates (23% to 25%). On May 9, Treatments 2 and 4 received an e-mail reminder. It appears that this helped and after May 9, Treatment 2 and especially Treatment 4 were in the lead, in terms of return rate, over the other two treatments. The advance was kept until the end of May for Treatment 2 and until mid-June

for Treatment 4. It seems that having a reminder action every two weeks instead of every month gave Treatments 2 and 4 the opportunity to advance at the beginning of follow-up period. On May 26 (E3), all four treatments received an e-mail reminder. On June 7 (E4), Treatment 2 received its fourth e-mail reminder and Treatment 4 its third one. However, this fourth e-mail reminder for Treatment 2 did not seem to have a similar positive impact as the three previous ones.

By June 20 (E5, the end of the experiment for Treatments 2 and 4), return rates were around the 50% mark for all four treatments. We were able to get more than 40% of questionnaires returned by just sending e-mail reminders and around 50% by just doing one phone attempt plus e-mail reminders. It is even more of interest to note that Treatment 4, where only one phone attempt was made between April 26 and April 29, attained almost the same results as Treatment 1 (standard treatment) and as Treatment 3, where telephone follow-up was applied continuously since April 26.

After June, Treatments 1 and 3 began to overtake Treatments 2 and 4. On August 1 (E8), these treatments exceeded Treatments 2 and 4 by more than 10%. Since regular telephone follow-up for Treatments 2 and 4 began after the end of the experiment (in fact on July 8 due to operational issues), the impact of the telephone follow-up was seen only at the end of summer. By the end of September, the four NRFU treatments show very similar return rates. This steady pattern continued until the end of collection with all treatments finishing with over 80% return rate.

Graph 1: Cumulative unweighted return rate observed on each day during collection

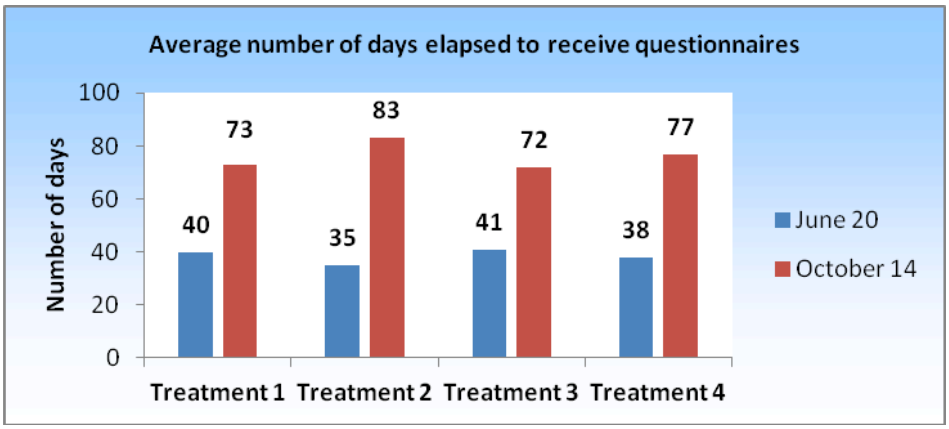


Once we looked at the seven surveys separately, we observed more variability among the treatments in the unweighted return rates. For the Head Office and the Specialized Design surveys, the rates for Treatments 1 and 3 were higher than those of the two other treatments. For the Software Development and Computer Services survey, Treatments 2 and 4 seemed to give the highest rates. Treatment 2 also gave the highest rate for the Accounting survey. For the Engineering survey, Treatments 1, 2 and 4 had equal rates but the rate for Treatment 3 was lower. For the Architecture survey, Treatment 3 was the highest. Although some differences in the return rates were observed among the surveys, at the end of collection, the return rate for each treatment in each survey was always 74% and above. In some cases the percentage was over 80% and it sometimes reached 90%.

4.1.2. Return Time

Graph 2 shows the average number of days elapsed to receive questionnaires for each treatment. The number of days elapsed is derived by the difference between the date when the questionnaire is submitted to Statistics Canada, and the EQ mail-out date of March 28. On June 20, Treatment 2 units had the shortest return time; however, all treatments were very close, with 6 days being the maximum return time difference between treatments. At the end of collection, on October 14, Treatments 2 and 4, in general, had a higher average number of days elapsed to receive questionnaires. Note that telephone follow-up for Treatments 2 and 4 started on July 8, one month after the last e-mail reminder was sent. This could lead to a longer average response time for these two treatments.

Graph 2: Average number of days elapsed to receive questionnaires for each treatment (on June 20 and on October 14, 2011)



4.1.3. NRFU Costs

The total cost of collection for UES surveys includes pre-contact, mail-out of questionnaires, NRFU, FEFU, capture and imaging, management and other tasks such as helpline, quality control monitoring, system testing, etc. In this paper, given the focus on NRFU strategies, we will concentrate on NRFU costs defined through the number of NRFU attempts and the length of such attempts among the four treatments.

A NRFU phone attempt is recorded in the Blaise System when an interviewer tries to contact the respondent at the time when a questionnaire has not yet been submitted to Statistics Canada. The start and end times of each attempt are recorded along with the outcome of the attempt. At each attempt, one or more of the following outcomes could result:

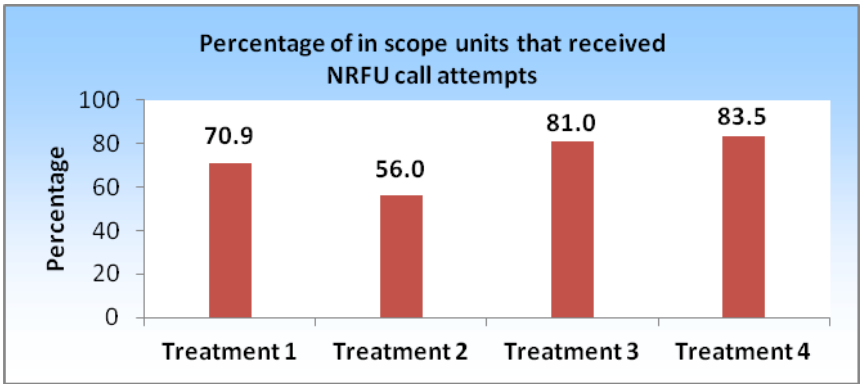
(1) An interviewer spoke with a primary respondent
(2) An interviewer spoke with an alternate contact
(3) An interviewer spoke with a secretary or a receptionist
(4) An interviewer left a message on the answering machine
(5) No one answered the call
(6) The line was busy
(7) Other outcomes

The percentage of in-scope units having at least one NRFU call attempt by the end of collection (October 14) for each treatment is presented in Graph 3. As seen from Graph 1, approximately 15% of the units in each treatment returned their questionnaire by April 26. Only the remaining outstanding units were eligible for NRFU.

Treatment 2 had the smallest proportion of units getting the NRFU attempts because outstanding units in Treatment 2 received only e-mail reminders in the first three months of the experiment. Telephone follow-up did not start until July 8 for this treatment. Note that some units in Treatment 2 recorded a call during the experiment because it was agreed at the beginning that if an enterprise called the help line at Statistics Canada, interviewers would call them back as needed.

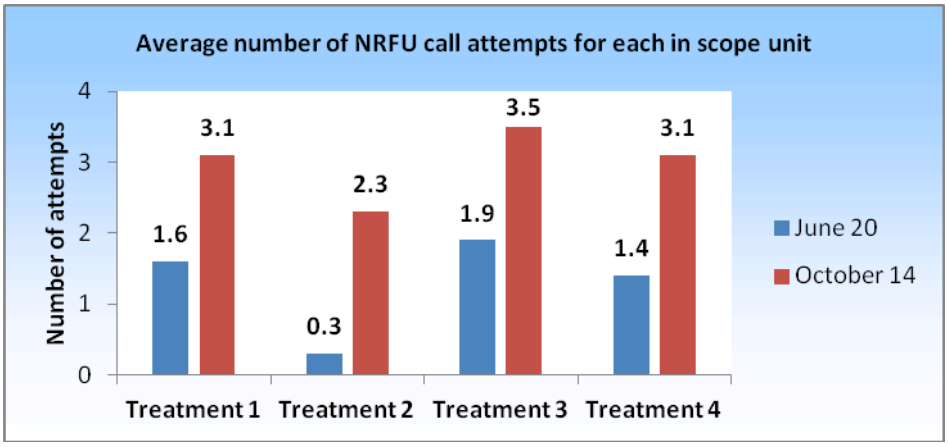
Treatments 3 and 4 had the highest number of units with at least one phone attempt due to the set up of the experiment: outstanding units in these two treatments received a phone attempt during the week of April 26. For Treatment 1, the outstanding units did not receive any telephone follow-up until after receiving the first e-mail reminder. Thus, the percentage of Treatment 1 units with phone attempts was relatively lower than for Treatments with the initial follow-up attempt done over the phone.

Graph 3: Percentage of in scope units that received at least one NRFU call attempt by the end of collection, October 14, 2011



The average number of call attempts for each in scope unit in the four treatments as of June 20 and October 14 are presented in Graph 4. Recall that a maximum of five NRFU phone attempts was allowed before a unit was considered as a final non-response unit. Recall also that not all units were eligible for NRFU (as they might have submitted their questionnaire by the time NRFU started) and even if they were, not all outstanding units were contacted by phone due to set up of the experiment. As of June 20, Treatment 2 was the only treatment that received e-mail reminders only. It is therefore expected to observe a very low average attempts for Treatment 2. The average started to pick up after July 8 since all outstanding units in the four treatments received telephone follow-up. By the end of collection, the average number of attempts for Treatment 2 was still the lowest, while for the other three treatments, the average were similar. As expected, Treatment 3 had the highest average number of NRFU attempts as it included both phone follow-up throughout collection as well as the initial phone follow-up attempt on April 26. It would also appear that effect of doing continuous telephone follow-ups, as was done in Treatment 1, and doing only the first follow-up attempt over the phone, as was done in Treatment 4, resulted in similar average number of NRFU call attempts.

Graph 4: Average number of NRFU call attempts for each in scope unit in each treatment (on June 20 and on October 14, 2011)

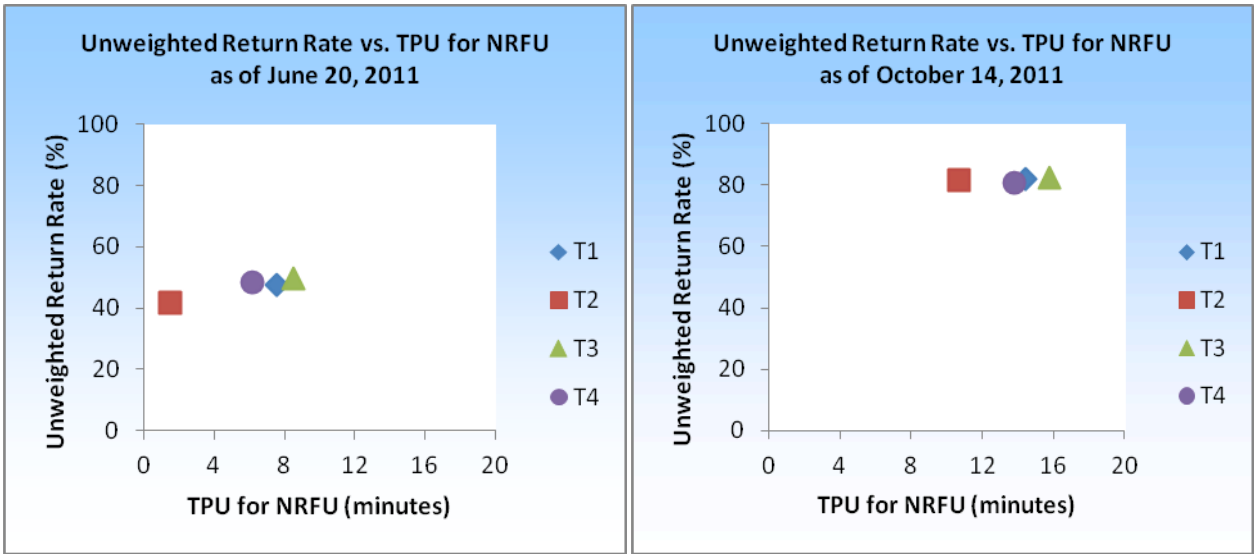


As mentioned before, the start and end times for each attempt are recorded in the Blaise System. We are then able to compute the duration of the attempts, also known as Time Per Unit (TPU). However, during an attempt, an interviewer sometimes has to consult his/her supervisor while the case is still open in the system. This increases the duration of the call recorded on Blaise, but the time recorded does not reflect the actual time spent on the telephone to reach one or more of the seven outcomes listed earlier. To reduce the impact of these unwanted prolonged attempts, each attempt is subject to an outlier detection strategy.

Since at each attempt there could be more than one outcome, we had to assign a dominant outcome. When an interviewer successfully spoke with a primary respondent or with an alternate contact during an attempt, it is considered a direct contact. If no direct contact was accomplished at a given contact attempt, it is considered an indirect contact. We then applied the following upper bounds to each attempt: 20 minutes for direct contacts and 15 minutes for indirect contacts. Attempts with duration exceeding the boundary values were subject to imputation. Imputation classes were identified for each combination of survey, treatment and contact (direct, indirect). Outlier attempts were imputed by the average values in each imputation class. We then used imputed values to compute the total duration of the attempts for all units, and divided this total duration by the total number of in scope units to get the average duration of the attempts per unit, i.e., the TPU.

One of the goals of this experiment was to find the most efficient non-response follow-up strategy, i.e., the one producing the best return rate at the lowest cost. Based on descriptive statistics, Treatments 1 and 3 were the treatments with the most intensive follow-up: they received e-mail reminders and telephone follow-up since the end of April. Consequently, they were also the most expensive. Treatment 4 was less expensive than Treatments 1 and 3 because there was no telephone follow-up until July 8, but there was one phone attempt at the beginning of NRFU. The least expensive strategy was Treatment 2, where only e-mail reminders were used at the beginning and there were no telephone attempts or follow-up until July 8. The scatter plots of the unweighted return rate versus the TPU for non-response follow-up on June 20 and October 14 are shown in Graph 5. Treatment 2 (red squares) had similar return rates as the other three treatments, but with a lower TPU spent on NRFU. Since it is more costly to spend time on telephone follow-up than to send e-mail reminders (which is an automated process), a treatment was considered more costly if the TPU on NRFU is higher. Therefore, from the scatter plots, we observed that Treatment 2 was able to give similar return rates with lower cost on telephone follow-up, compared to Treatments 1, 3 and 4.

Graph 5: Scatter plots of unweighted return rate versus the average duration of NRFU attempts (TPU) (on June 20 and on October 14, 2011)



4.2. Analysis of Variance

Analysis of variance was performed on weighted and unweighted return rates, number of days elapsed to receive questionnaire, the number of NRFU attempts and time spent per unit (TPU) on NRFU.

Recall that the Head Office Survey is a census where each unit has the same sampling weight, while for Services surveys a stratified sample is selected. Because of the difference in the sampling designs, the analysis of variance tests were run separately for Head Office and for Services surveys. For tests on Head Office units, the F-Test in the SAS procedure PROC GLM was used. For Services surveys, the Wald Test adjusted for design (Van den Brakel and Renssen (2005)) was used.

In Table 1 below, the results of analysis of variance tests showed that on both June 20 and October 14 (end of collection), the unweighted and weighted return rates were not significantly different among the four treatments for all surveys. The number of NRFU attempts and NRFU TPU of the four treatments were significantly different. The number of days elapsed to receive questionnaire was significantly different among the four treatments for the Head Office Survey as of June 20, but was not significantly different on October 14. It was not significantly different for Services surveys on both June 20 and October 14.

Table 1: Testing treatment differences on June 20 and on October 14, 2011

Main Effects: T1 = T2 = T3 = T4	Head Office		Services	
	Jun 20 p-value	Oct 14 p-value	Jun 20 p-value	Oct 14 p-value
Weighted return rate	—	—	0.9996	0.9999
Unweighted return rate	0.0525	0.1200	0.6295	0.9994
Number of days elapsed to receive questionnaire	<0.0001	0.7766	0.1038	0.5222
NRFU attempts	<0.0001	0.0228	<0.0001	0.0063
NRFU TPU	<0.0001	0.0389	<0.0001	0.0098

5. Conclusions

We successfully implemented an embedded experiment to test four strategies for the non-response follow-up of electronic questionnaires for seven UES surveys in the 2011 collection cycle. The design of experiment showed that it might not be worth starting telephone non-response follow-up during the first two or three months of collection. The strategy consisting of e-mail reminders every two weeks at the beginning of collection succeeded in obtaining the first 40% of response at lower cost. Results showed that even if telephone follow-up starts three months later than in the current follow-up strategy, very similar final response rates can be achieved, with fewer phone attempts and less effort.

The initial tests run to compare implemented treatments seemed to support these conclusions obtained from the descriptive analysis. The random allocation of the intervention or treatment helps to ensure no systematic differences between intervention groups in factors, known and unknown, that may affect outcome. However, more detailed analysis is yet to be performed to finalize the overall conclusions from this experiment. In particular, we would like to look at contrasts between the treatments to do statistical analysis of such factors as the first attempt done over the phone. We will also need to look at the implications of using different modes for the collection and assess if there is any impact on the estimates.

6. Recommendations

Statistics Canada will continue to use electronic questionnaires on these seven UES surveys in the next cycle. By 2014, all UES surveys will use electronic questionnaires as the primary mode of collection. The experiment described in this paper was aimed at finding the most efficient strategy of NRFU on electronic questionnaires to be used in the future UES cycles. From the results of the experiment, a NRFU strategy that consists of sending only e-mail reminders in the first months of collection, until the return rate reaches a plateau and then doing a regular telephone follow-up until the end of collection appears to be the most efficient. Further tests will be run to confirm this initial conclusion that would lead to recommendations for Statistics Canada to implement this strategy in 2012 for these seven surveys.

Acknowledgement

This experiment was made possible due to efforts of many areas within Statistics Canada. Authors would like to acknowledge all the hard work and contribution by Service Industries Division, Enterprise Statistics Division, Business Survey Methods Division, Social Survey Methods Division, Collection Systems and Infrastructure Division, Operations and Integration Division, Collection Planning and Management Division, and Sturgeon Falls Regional Office that made this study possible.

References

Evra, R.-C. and DeBlois, S. (2007). Using paradata to monitor and improve the collection process in annual business surveys. Proceedings of International Conferences on Establishment Surveys, ICES-III.

Groves, R.M. and Lyberg, L. (2010). Total survey error past, present, and future. *Public Opinion Quarterly*, 74(5), 849-879.

Jackle, A., Roberts, C. and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78(1), 3-20.

Van den Brakel, J.A. and Renssen, R.H. (1998). Design and Analysis of Experiments Embedded in Sample Surveys. *Journal of Official Statistics*, 14(3), 277-295.

Van den Brakel, J.A. and Renssen, R.H. (2005). Analysis of experiments embedded in complex sampling designs. *Survey Methodology*, 31(1), 23-40.

Van den Brakel, J.A. and Van Berkel, C.A.M. (2002). A design-based analysis procedure for two-treatment experiments embedded in sample surveys. An application in the Dutch labor force survey. *Journal of Official Statistics*, 18(2), 217-231.