UNDERSTANDING, PREDICTING AND MITIGATING WEB SURVEY BREAKOFFS

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Humanities

2022

Zeming Chen

School of Social Sciences, Department of Social Statistics

Table of Contents

WORD COUNT: 43,843

List of Tables

List of Figures

Abstract

Web survey respondents quit survey partway through more frequently than in other survey modes. This pre-mature quitting event is called survey breakoff. It causes missing data, reduces sample size, lowers statistical power, and sometimes biases survey estimates. Using a number of experiments, statistical models and simulations, this thesis contributes to the understanding, prediction and mitigation of web survey breakoffs. It tackles breakoffs from three stages of the survey data collection: before, during and after the survey.

Chapter 4 focuses on the survey design stage by randomly allocating survey respondents to one of the filter question formats and one of the six orders of the question topics. It shows that presenting all filter questions before showing any follow-ups (i.e. grouped filter question format) postpones the breakoff, compared to presenting them by pairs (interleafed format). However, as respondents answer more questions, the breakoff rate in the grouped format quickly catches up with that in the interleafed format. Additionally, when introducing upcoming new topics, more breakoffs are expected. Meanwhile, the insurance-related topic has more breakoffs than the topics about clothing purchase and utilities payment.

Chapter 5 predicts breakoff during the survey using seven statistical models (traditional and LASSO Cox, traditional and LASSO logistic regression, Support Vector Machine, Random forest, and Gradient boosting) and four types of predictors: (1) respondents' demographics, (2) time-varying variables (whose values change by questions) coded concurrently, (3) time-varying variables coded cumulatively, and (4) the three previous predictors together. The gradient boosting produces the best performance for breakoff prediction while the Cox survival model does not improve the prediction further although it accounts for the clustered structure in the breakoff data (questions clustered within respondents). Also, time-varying variables are best used concurrently to improve the prediction of breakoff.

Chapter 6 investigates different strategies for adjusting for breakoff after the survey. Four methods are applied to the simulated data where four breakoff rates and three breakoff mechanisms are manipulated, and their ability to mitigate the breakoff bias is compared. It is found that multiple imputation outperforms the other three methods employed in the study and the cause of breakoff is more influential on the effectiveness of compensation methods compared to the breakoff rate.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

"Why do you want to do research in survey methodology?" I was asked this question during the interview when applying to the PhD programme. My answer to it was very simple: I have interest in it. After the interview, I blamed myself for giving such a clichéd and boring answer. Back then, I thought the "ideal" answer should be something extraordinary, such as the ambition to push forward the boundary of human beings' knowledge.

I was wrong. Looking back on the past three years, having an interest in the topic is actually an excellent answer to the question. Without an interest and curiosity in the research topics, I might never have been able to go through the numerous setbacks in the research, remain motivated, and see the end of this PhD.

Thanks to my supervisors, I had many chances to explore and pursue my interest over the past three years. Both my supervisors Alexandru Cernat and Natalie Shlomo surely offered help in various aspects of the research, such as data cleaning and concise writing. However, I think what needs more acknowledgment is their constant encouragement for me to explore my interest, both in academia as well as personal career aspirations. They gave me total freedom and provided timely and constructive guidance along the way. Because of this, I could try different things without fear of any possible consequences. I am incredibly grateful for this as well as their professional guidance over the past three years.

If completing a PhD needs two resources, then having an interest is just the internal resource. The external resource is also important. Firstly, the scholarship from the School of Social Sciences gives me a peace of mind while I am conducting the research. Also, I am thankful for the opportunities to discussing my research in the departmental research seminars as well as meeting other researchers such as Stephane Eckman, Joe Sakshaug, and Eduardo Fe. All of them provided invaluable feedback and inspiration for my research. I also want to thank my PhD fellows: Hafsteinn Einarsson, Jiyao Sun, Ting Liu, and Xinyi Kou. It is through many informal conversations I had with them that I realise the wider impact of my research. Also, the importance of their emotional support can never be emphasised enough.

Finally, none of this would have been possible without the support of my parents. I want to dedicate this PhD thesis to them.

Chapter 1 Introduction

Surveys are widely used in society. Governments use surveys to gather the public's opinions and to inform and evaluate their policies (Department for Digital, Culture, Media & Sport, 2022). Companies rely on surveys to understand customers' perceptions about their products (Muggah and McSweeney, 2017). Researchers collect data via surveys to generalise findings to the target population (Bekova, 2021). Since the turn of the century, web surveys have increasingly been conducted to achieve all these goals. This is mainly because this specific survey mode has a lower cost and a shorter turnaround time compared to interviewer-administered modes such as face-to-face surveys (Groves, 2011).

1.1 The representativeness of web surveys

However, web surveys also come with some disadvantages. Its main challenge is related to the representativeness of the collected sample. As some individuals do not have the Internet access, conducting surveys online will exclude these individuals from the sample and potentially result in the coverage bias (Couper, 2000). Another related issue is self-selection. There are two types of web surveys (probability and non-probability), and many current web surveys are non-probability and opt-in based. Participating in the opt-in survey will require respondents to see the online survey invitation and self-select to cooperate with the invitation (Callegaro, Lozar Manfreda and Vehovar, 2015). In this case, some individuals who are not online during the survey period or who are online but do not see the survey invitation will have a zero probability of being recruited into the survey. Both under-coverage and self-selection violate one of the prerequisites for producing unbiased survey estimates. That is, every individual in the population must have a known and non-zero probability of selection (Bethlehem, 2010). As a result, the findings from the non-probability web surveys become less applicable to the general population. To make inference to the general population, the probability-based web survey is usually recommended (Cornesse *et al.*, 2020), and the push-to-web approach is commonly adopted when recruiting the sample in this type of survey. In such an approach, sample members are randomly drawn from the sampling frame with known and non-zero probabilities and then invited to respond to the web survey using means other than the Internet (e.g., letters, telephone calls, personal visits) (Scherpenzeel, 2011; Bosnjak *et al.*, 2018).

Even though no bias exists in the sample recruitment stage, the representativeness issue can still be present in the probability-based web surveys, especially when the survey nonresponse occurs. Compared to other survey modes, web surveys tend to have a lower response rate (Daikeler, Bosnjak and Manfreda, 2020). A low response rate reduces the available sample size and risks introducing biases to survey estimates when respondents differ from nonrespondents (Bodor, 2012).

The representativeness issue related to survey nonresponse has been studied extensively, and many efforts have been made to increase the response rate, such as increasing the response rate by offering incentives (Noel and Huang, 2019) or allowing the sample members to choose their preferred survey modes (Olson, Smyth and Wood, 2012). Other studies have also been conducted to use statistical methods to compensate for the nonresponse bias (Biemer and Christ, 2008; Bonander *et al.*, 2019). All these measures aim to ensure the representativeness of the final collected survey data.

1.2 The problem of survey breakoffs

No matter how effective those measures are, researchers still cannot rest assured that their collected data are well representative of the target population. In fact, survey response can be viewed as a sequence of decisions (Mittereder, 2022): the invited sample member first decides to visit the website of the questionnaire, glances through the study introduction page and then determines whether to answer the survey questions. At each survey question, respondents repeatedly go through the process, deciding whether they should continue or quit. Given that there are usually multiple questions in the survey, the response burden can overwhelm the respondents and ultimately discourage them from completing the entire survey questionnaire. When people access the survey but do not complete it, survey breakoff happens (Lavrakas, 2008). This event is sometimes referred to as survey dropout, incompletion or partial response (Mittereder, 2022).

Survey breakoff can bias the final collected data. This is true no matter the sampling is probablity or non-probability based. For the probability surveys, if the respondents who break off from the survey differ from those who complete the survey, the analysis result has a risk of being biased (Steinbrecher, Roßmann and Blumenstiel, 2015). When the survey sampling is non-probability such as the quota sampling commonly used in the randomised clinical trial, participants with certain characteristics might break off more often in the trial (Leon *et al.*, 2006). This can damage the causal inference from the trial. Therefore, tackling breakoffs is equally important for probability as well as non-probability surveys, and this thesis is devoted to tackling the issue of survey breakoff in both types of surveys.

Currently, the survey breakoff rate is not commonly reported by survey agencies, but some researchers have collated the breakoff rates of different surveys. For example, after analysing the breakoff rate across 186 non-probability web surveys, Revilla (2017) noted that the breakoff rate is wide-ranging (1.1% to 62.1%) with an avarage of 12%-13%). Considering that many existing well-known surveys have a large number of sample members, a 10% dropout can easily amount to hundreds of cases lost. Not being able to collect all the desired information from hundreds of breakoff cases means that part of the money and efforts spent on recruiting is wasted. Furthermore, survey breakoffs will lead to missing data in the collected data, which means that users have a smaller number of cases and lower statistical power in their analysis.

1.3 Existing research gaps

Given the negative influence of survey breakoff, the topic has received considerable attention. All existing literature focused on three aspects of breakoff, each of which corresponds to one stage of survey data collection. The first stage is **before** the survey starts. The research focusing on this stage investigates what survey design factors are related to the survey breakoff event. Some of the examples are the use of the progress bar (Matzat, Snijders and van der Horst, 2009) and the length of the question (Tijdens, 2014). The second type of literature concentrates on tackling breakoff **during** the survey. To be specific, researchers create statistical models to predict respondents' breakoff likelihood while they are responding to the survey. For respondents who are predicted to break off soon, the model will trigger interventions to keep them engaged (Mittereder and West, 2021). The third way of tackling the negative effect of breakoff happens **after** the survey data collection. At this stage, statistical methods are used to compensate for the missing data caused by breakoff. An example of this is breakoff weighting, which increases/decreases the impact of certain observations on the survey estimates if they are under/over-represented in the final survey data (Steinbrecher, Roßmann and Blumenstiel, 2015).

Thus, the existing literature tackles the breakoff issue from three different perspectives: (1) **understanding** the impacting factors of breakoff and optimising the design **before** the survey starts, (2) **predicting** the imminent breakoff event and intervening **during** the survey, and (3) **mitigating** the impact of breakoff **after** the survey data collection. This thesis follows the past literature and further develops the research in these three aspects.

Regarding the first stage (i.e. the design of the survey), the design of filter questions and the topic of the questions receive little attention so far. Many surveys have filter questions, which can trigger follow-ups when respondents choose "yes" as the answer. When seeing this type of questions, respondents can easily learn about the extra response burden and decide to quit the survey. Different ways of presenting the filter question and its follow-ups are available, but few researchers investigated how those designs can influence breakoff. Meanwhile, question topics are also influential on respondents' breakoff tendency. This is particularly true when the topics are uninteresting to the respondents (Shropshire, Hawdon and Witte, 2009). Currently, the knowledge about the impact of question topics on breakoffs is limited because prior published research has not randomised the order of the topics. Previous research has shown that the amount of time spent in the survey and the order of the topic can influence respondents' tendency to respond to the next question (Teclaw, Price and Osatuke, 2012; McGonagle, 2013). Therefore, the lack of random allocation of topic orders can lead to biased results in past literature. This is because it is unclear whether the question order, the topic itself, or the response burden accumulated from the beginning of the questionnaire is affecting the breakoff event.

In addition to studying the impact of the filter questions and question topics using a randomised design, it is also important to consider the breakoff timing (how many questions respondents have seen/answered prior to breakoff). The breakoff timing can give valuable information regarding the design of the study as even with the same amount of breakoff survey designers will prefer respondents to answer more questions.

During survey data collection (the second stage discussed above), the effectiveness of real-time interventions depends partially on good prediction models. Ideally, the model should generate an accurate prediction of breakoff propensity so that the intervention can be triggered at the most relevant time. However, only the traditional Cox survival model is used as the predictive model in the existing literature. Meanwhile, machine learning models are increasingly popular in the research community but have not been applied in the context of breakoff prediction. It is, therefore, unclear what statistical models are more predictive of the survey breakoff. Also, given that survey response consists of a sequence of respondents' actions at each question, researchers have repeated measurements of these information (i.e. time-varying predictors). It is unknown from the current literature how to best code these kinds of variables to maximise prediction performance.

The final stage (i.e. after survey data collection) currently receives the least attention among all three stages. Considering that breakoff is a special type of survey nonresponse, the techniques to correct for the missing data caused by the survey nonresponse (weighting, imputation, etc.) can be applied to the breakoff problem. Nonetheless, it is unknown from past studies whether those techniques will perform as expected in the context of breakoff, especially considering that breakoff happens at the question level while some of those techniques work at the unit/respondent level (e.g., weighting). Another related question is how the breakoff should be treated when adjusting for it. Researchers could choose to combine it with unit nonresponse or treat it as a separate nonresponse process. By definition, breakoff can be considered as a survey (non)response, so there may be no need for a separate compensation for it. On the other hand, even though breakoff and unit nonresponse have some impacting factors in common, previous research noted that breakoff has its own impacting factors as well (Peytchev, 2009). Therefore, it might be best to treat breakoff as a unique survey outcome and correct for it separately. A limited amount of literature has empirically investigated the implication of accounting for breakoff separately.

1.4 Rationale for the alternative format

The present thesis will contribute to all three research stages highlighted above using the alternative/journal format. This format is chosen mainly because of the workflow in this PhD. To be specific, the research of every year is focused only on one of the three stages of survey breakoffs (i.e. before, during, and after survey data collection). At the end of each year, the research result is submitted to the peer-reviewed journal. The alternative format fits well with this workflow as it allows me to incorporate chapters that are already or will be submitted to peer-reviewed journals.

Another reason for choosing the alternative format is that the present work constitutes a body of publication tending towards a coherent and continuous thesis. To be specific, although Chapter 4, 5 and 6 in this thesis are three standalone research papers that approach the survey breakoffs from different aspects, they focus on the same overarching research topic: tackling web survey breakoffs.

1.5 The structure of this thesis

The three substantive chapters described above and other chapters are organised in this thesis as follows. A literature review is first presented in Chapter 2 to give an overview of the current state of art in the study of survey breakoff. It will also provide a more detailed account of the existing research gaps and justify the necessity of filling them. Following this, the data and analysis methods used in this thesis will be discussed in Chapter 3. Then, three substantive chapters (Chapter 4 to 6) are presented, which focus on understanding, predicting and mitigating breakoffs, respectively.

Chapter 4 concentrates on the survey design stage and investigates how the filter question formats and question topics affect the breakoff event and its timing. This chapter uses experimental survey data where respondents were randomly allocated to one of the two filter question formats and one of the six question topic orders. Variables about the two experiment designs and their interaction with time (represented by the number of questions respondents saw) will be included in a series of Cox survival models along with other control variables. This model specification and experimental design will contribute to the understanding of how these two design decisions impact breakoffs.

Chapter 5 tackles the breakoff issue during the survey response by comparing different statistical models and predictors. In total, seven statistical models (including machine learning techniques) are fitted along with four different types of predictors. These seven models are: traditional and LASSO Cox survival model, traditional and LASSO logistic regression, Random forest, Gradient boosting, and Support Vector Machine. The four types of predictors are (1) respondents' demographic information only, (2) time-varying predictors only (coded concurrently), (3) time-varying predictors only (coded cumulatively), and (4) the three previous predictors together. In the end, 28 models (7 model types × 4 predictor types) are compared based on their performance in breakoff prediction. Their prediction performance is measured by six metrics: C-index, Accuracy, Sensitivity, Specificity, Precision and AUC. By comparing these metrics across both models and predictor types, Chapter 5 will contribute to the real-time breakoff interventions by proposing the best combination of predictive model and the coding of predictors.

Chapter 6 shifts the focus to the post-survey adjustment and investigates the methods for mitigating the impact of breakoff. It is a simulation study where the response data from a cross-national probability web survey is used as the base to create the simulated population. Based on this population, survey nonresponse and breakoff are separately simulated. When simulating breakoff, its rate and cause are manipulated. Its rate changes from 5% to 20% at an increment of 5%. Three causes of the breakoff are tested as well: (1) being completely random (i.e. Missing Completely At Random), (2) being impacted by only observed variables (Missing At Random), and (3) being influenced by both observed as well as unobserved variables (Missing Not At Random). In total, 12 breakoff scenarios are created (4 breakoff rates × 3 missing data mechanisms). To deal with the breakoff and its resultant missing data, four different methods are tested: (1) completely ignoring breakoff, (2) classifying breakoff as survey nonresponse and using only the nonresponse weighting in the data analysis, (3) treating breakoff as a separate outcome of survey nonresponse and weighting the data by a combined nonresponse and breakoff propensity, and (4) multiple imputation. While the first two methods do not distinguish breakoff from survey nonresponse, the latter two do. These four methods are applied to the 12 breakoff scenarios to estimate two statistics of interest: univariate means, and model coefficients in multivariate analysis. The estimates from these four methods are compared with the benchmark values obtained from the simulated population. By empirically investigating the effectiveness of different breakoff compensation methods under different breakoff rates and patterns, Chapter 6 will contribute to the discussion regarding the most appropriate method to correct for breakoff.

This thesis will end with Chapter 7 where the findings across the three substantive chapters will be tied together to give recommendations for designing the survey to be considerate of breakoffs from its beginning to the end. The contributions of this thesis will also be highlighted along with the limitations and future research opportunities.

The three substantive chapters described above have been co-authored, and two of them have been published in peer-review journals. Here I present the publications and contributions of different authors.

Chapter 4 was co-authored with Alexandru Cernat, Natalie Shlomo and Stephanie Eckman. I have designed all the research stages, carried out the data cleaning, data manipulation and data analysis as well as written the draft for all the sections of the paper. Stephanie provided access to the data. Alexandru gave guidance on how to clean the data. All three co-authors gave feedback on the draft paper, and we worked together when revising the paper. This chapter has been published as:

- Chen, Z. *et al.* (2022) 'Impact of question topics and filter question formats on web survey breakoffs', *International Journal of Market Research*. doi: 10.1177/14707853211068008.

Chapter 5 was co-authored with Alexandru Cernat and Natalie Shlomo. I was responsible for designing all the research stages, cleaning, manipulating and analysing the data. In addition, I wrote all the sections of this paper. Alexandru, Natalie and I worked together in revising the paper. This chapter has been published as:

- Chen, Z., Cernat, A. and Shlomo, N. (2022) 'Predicting web survey breakoffs using machine learning models', *Social Science Computer Review*. doi: 10.1177/08944393221112000.

Chapter 6 was a joint work between Alexandru Cernat, Natalie Shlomo and me. I designed all the research stages, simulated and analysed the data as well as wrote the draft for all the sections. Alexandru and Natalie guided me through the data simulation process, and we worked together in revising this chapter. This chapter will be submitted to the peer-reviewed Journal of Survey Statistics and Methodology.

Chapter 2 Literature Review

2.1 Prevalence and consequence of survey breakoffs

Survey breakoff is prevalent, and its extent varies dramatically according to Revilla (2017) who reviewed 185 non-probability surveys. While some surveys barely had any breakoff events (1.1% breakoff rate) others suffered greatly from it (with a breakoff rate of 62.1%). After aggregating the data, Revilla (2017) reported an average breakoff rate of 11.8%. A different meta-analysis, which was conducted by Liu and Wronski (2018), documented a similar average breakoff rate. Across 25,000 non-probability web surveys in their study, they found an average breakoff rate of 13% with a standard deviation of 10%. Hoerger (2010) not only reported the wide-ranging breakoff rate across six student surveys in his study (6% to 30% breakoff rate) but also quantified the relationship between breakoff rate and survey length (measured by the number of survey questions) using a linear regression. The model result showed that approximately 10% of participants would break off immediately after the survey began and additional 2% of participations were expected to break off after every 100 questions.

The prevalence of breakoff means that missing data are common in web surveys. As a result, users of survey data will have fewer observations in their analysis, which leads to smaller statistical power and larger variations in survey estimates. This can also be problematic for the comparison of nested models. As the model might include independent variables that suffer from different missing data patterns due to breakoff, models can be based on different samples. This makes it difficult to compare models unless some solutions are adopted (complete case analysis, imputation, etc).

The worst consequence of breakoff appears when survey estimates are biased because respondents who do not complete the survey are different from those that do. Indeed, the presence of breakoff bias is documented in the McCoy *et al.* (2009) study. They conducted a web-based probability survey among students of 10 U.S. colleges and asked about their drinking and smoking behaviours. The comparison between the complete respondents and the breakoff cases showed that students who engaged in high-risk drinking behaviours were more likely to quit the survey. As a consequence, the survey estimates about drinking (e.g., number of days being drunk in a typical week) were biased downwardly.

2.2 Factors impacting survey breakoffs

Considering the negative influence of breakoff, many studies have been conducted to investigate factors that impact breakoffs. All those factors can be classified into four categories (Peytchev, 2009; Mittereder and West, 2021). The first category refers to the survey design, and it mainly concerns design decisions that impact the entire survey. Some examples are the use of incentives (Silber, Lischewski and Leibold, 2013), the number of questions in the survey (Hoerger, 2010), and modularisation of the survey (i.e. surveys are split into parts and participants respond to the parts consecutively according to a fixed interval over a period of time) (Toepoel and Lugtig, 2018).

The second category of factors related to breakoff refers to the characteristics of survey pages and questions. These features can be seen by survey participants only after they start the questionnaire. Some examples of question characteristics that have been found to be associated with breakoffs are open-ended question and the number of characters in the questions (Peytchev, 2009; Tijdens, 2014). The way these factors affect breakoff can be explained by the survey response theory. According to Tourangeau (2018), when responding to a survey question, respondents have to go through a series of steps, ranging from understanding the questions, retrieving the required information from the memory to mapping the information to the provided answers. Questions that are cognitively demanding will likely cause burden in any of the steps, which can subsequently discourage respondents from continuing the survey. An empirical study that highlights the effect of survey burden on breakoff can be found in Galesic (2006). In her study, breakoff cases self-reported a higher level of burden before they quit the survey.

Respondents' socio-demographic information is the third category of factors influencing breakoffs. They are usually used as a proxy for respondents' ability to deal with the response burden or their general tendency to cooperate with the survey request. The findings in this research area are mixed. However, the overall conclusion is that respondents who have the following characteristics are more likely to break off: older, less educated, male, non-white, student and affluent (Galesic, 2006; Peytchev, 2009; Klein *et al.*, 2011; Mittereder and West, 2021).

Paradata, the final category of factors affecting breakoff, refer to the by-product information collected during the response process (McClain *et al.*, 2018). Some examples are question response time (Zhang and Conrad, 2014), mouse movement (Fernández-Fontelo *et al.*, 2021) and answer changes (Stern, 2008). Paradata are believed to reflect the respondents' changing motivation and response burden (Mittereder and West, 2021). For instance, Horwitz, Kreuter and Conrad (2017) set up a lab experiment where questions with easy-to-understand and complex wordings were randomly presented to participants. After answering each question, participants were asked about their perceived difficulty of the question on a five-point scale. Whilst participants were responding to the survey questions, their mouse movements were also recorded. In the end, Horwitz and her colleagues found that there were more mouse movements in the question whose wordings were perceived to be complex.

2.3 Three ways of tackling survey breakoffs

Based on the factors identified in the four categories above, survey researchers have proposed different methods to tackle survey breakoffs. All those studies can be classified into three strands based on the timing when the breakoff issue is tackled: before, during and after the survey.

Studies in the first strand focus on the survey design stage (i.e. before the survey starts) by proposing user-friendly designs. The literature of this research strand usually uses experiments where a specific design element is manipulated and the resultant breakoff rate is compared between the control and experimental groups. Once the manipulated element is found to be associated with differential breakoff rates between the groups, relevant design recommendations are made. For instance, in an experiment conducted by Conrad *et al.*(2010), there were four designs of progress bars (which visualised the progress respondents had made in the questionnaire) and respondents were assigned to one of them at random. The four conditions were (1) no progress bar (i.e. control group), (2) linear progress bar (the bar moved linearly), (3) fast-to-slow progress bar and (4) slow-to-fast progress bar. The result showed that participants seeing the slow-to-fast progress bar broke off more often compared to other three groups. Based on these findings, Conrad and his colleagues discouraged the use of slow-to-fast progress bar. Sischka *et al.* (2022) also carried out an experiment but about the force-answering (i.e. respondents could not proceed to the next question unless answering the current one). Participants of their experiment were assigned to one of the groups at random: forced-answering vs. non-forced-answering. The fact that the former group suffered from more breakoffs led to their recommendation against the forced-answering design.

In addition to implementing experiments to justify survey designs, researchers also developed statistical models to examine whether the relationship between the design of interest and breakoffs is statistically significant and then make suggestions accordingly. For instance, by including the number of characters in the question as an explanatory variable in the model for survey breakoff, Tijdens (2014) found a statistically significant relationship between this variable and breakoff. As a result, she advocated for shorter questions in the survey. In another example, Peytchev (2009) included many variables about different characteristics of questions in a model to explain breakoffs and found that cognitively demanding questions (e.g., open-ended questions) and technically complicated questions (e.g., questions with a slider bar) were associated with more breakoffs. Based on the findings, he called for a careful consideration on the use of these types of questions.

As can be seen, studies in the first strand of research on breakoff followed the path of identifying the design associated with breakoff (via experiment or modelling) and then proposing some optimised survey designs to facilitate the survey completion. However, no matter how optimised the designs are, they cannot eliminate all the response burden. As discussed previously, answering a survey question can involve considerable mental processing, so the more questions answered will inevitably lead to a growth in survey burden, which can subsequently increase respondents' breakoff likelihood. As reviewed earlier, the response burden can be reflected in paradata. In this case, a question naturally arises: can we use paradata captured during the survey to predict respondents' breakoff likelihood and then intervene in real time (instead of relying on the reactive approach to optimising survey designs)?

The second way to tackle the breakoff is through proactive interventions during the survey response process. More specifically, survey designers can use paradata (along with the three categories of factors mentioned above) in a statistical model to continuously predict the breakoff propensity for each respondent at each question. When the predicted breakoff propensity exceeds a pre-defined threshold, the model can trigger interventions to encourage respondents to remain motivated.

Mittereder and West (2021) present an implementation of such a real-time intervention system. They included many variables (including paradata such as item nonresponse rate, and change in the question response time) in a logistic regression model to predict respondents' breakoff propensity. After implementing this model in their survey, they randomly assigned survey participants to one of the three groups in the experiment about the intervention. The intervention in their experiment was a pop-up message that praised the efforts respondents had put into the response and encouraged them to continue. While the control group never experienced a pop-up message (i.e. no intervention), respondents in another group (called generic group in their study) saw the message at the first question (i.e. intervening regardless of the predicted breakoff risk). In the third group (called tailored group in their study), the message was triggered when the model predicted that the respondents would break off at the next question (i.e. intervening at the highest breakoff risk). In the end, Mittereder and West (2021) noted that intervening in real time, compared to no intervention at all, could reduce the breakoff rate among respondents with certain characteristics such as students and females.

The review so far proves that it is helpful to tackle the breakoff issue before the survey starts or while the survey response is still ongoing. However, no matter how effective both approaches are, they can hardly guarantee that all respondents will complete the questionnaire. Therefore, at the final stage (i.e. after the survey), methods to compensate for the breakoff are investigated. In the literature focusing on this stage, breakoff is considered as a special case of survey nonresponse. The techniques for addressing the survey nonresponse problem can therefore be applied to the context of breakoff.

One of those techniques is weighting. Weights have been commonly used to correct for survey nonresponse bias, and they are essentially a set of numeric values that increase/decrease the impact of under/over-represented respondents on the analysis (Toepoel, 2015). To derive such weighting, a statistical model (e.g., logistic regression, classification tree) is used where the survey response status is explained by some variables (e.g., age, gender) that are available for both respondents and nonrespondents. Following this, for every respondent in the data, the model generates the response propensity whose reciprocal becomes the weights (Buskirk and Kolenikov, 2015).

Although applying weighting to account for breakoff seems to be straightforward, there is surprisingly a lack of papers about such applications. Steinbrecher et al. (2015) study is one of the few relevant publications. In their paper, they followed up with those who broke off from the German Longitudinal Election Study and compared the proportion of undecided voters in three analysis scenarios. In the first scenario, only the respondents to the initial surveys were analysed (i.e. they pretended that no follow-up was conducted). In the second scenario, both initially complete respondents as well as follow-up complete respondents were used in the analysis. To account for breakoffs in the follow-up study, frequency weighting was applied to those follow-up complete respondents. In the third scenario, only initially complete respondents were included in the analysis but weighted by their breakoff propensities. A logistic regression was employed to estimate the breakoff propensity using survey participants' demographic information (e.g., gender, age), interest in politics and responding device. In the end, they concluded that the estimated proportion of people who were still undecided about their preferred party in the 2009 German federal election would increase if the data of the breakoff cases were to be accounted for in the analysis.

2.4 Research gaps

As discussed so far, researchers approached the breakoff issue by focusing on one of the three stages of the survey data collection: before, during and after the survey. Following this categorisation, this thesis will present three substantive chapters, each of which addresses the breakoff issue from those three stages respectively.

2.4.1 Before survey data collection

Chapter 4 focuses on understanding how design decisions made before the data collection can impact breakoffs. More specifically, this chapter investigates how question topics and filter question formats impact breakoff. As will be discussed below, existing studies either confounded the impacts of these two designs on breakoff with the effect of other factors or did not take into account the breakoff timing.

How question topics impact nonresponse has received considerable attention in the past (Groves, Presser and Dipko, 2004; Roster, Albaum and Smith, 2017). According to Shropshire, Hawdon and Witte (2009), topics that were uninteresting to the respondents suffered from more breakoffs. However, they did not use an experimental design, so the order of the topic was always fixed. Past research already noted that the order of the topic can influence respondents' tendency to respond to the next question (Teclaw, Price and Osatuke, 2012). Meanwhile, a sizeable number of breakoff events have been reported to happen early in the survey (Peytchev, 2009). Without randomising the order of question topics, past research confounded the influence of three elements: the question order, the question topic and the response burden that accumulates from the beginning of the questionnaire. As a consequence, existing literature may misrepresent the real impact of question topics on breakoff.

In addition to the question topic, there is also a gap in the existing literature about how the design of filter questions affects the breakoff timing. Filter questions trigger follow-ups if answered positively. For instance, if the respondent answers "yes" to the question "Did you buy a T-shirt over the past 12 months?", they will see some follow-ups such as "How much was this T-shirt?" and "Where did you buy it?". There are two main ways to present the filter questions and follow-ups. In the *grouped* format, all filter questions are asked first before any follow-up is displayed. In the *interleafed* format, follow-up questions appear immediately after its corresponding filter question. According to the past literature, both formats can trigger extra questions and cause response burden, so respondents answering the two formats are equally likely to quit the survey (Kreuter *et al.*, 2011; Eckman and Kreuter, 2018).

However, what is currently unclear in the literature is the difference in the breakoff timing. Researching the breakoff timing is as important as studying the binary breakoff event. This is because survey designers will ideally prefer respondents to answer more questions even if the final breakoff rate would be the same. In the grouped format, respondents can only learn about the extra burden after going through all filter questions whereas those answering the interleafed format learn about the extra burden after one or two pairs of the filter questions and follow-ups. Nonetheless, the impact of grouped and interleafed formats on breakoff timing has not been empirically investigated.

Chapter 4 will fill the two research gaps using an experiment. In this study, respondents were randomly assigned to one of the two filter question formats and one of the six question topic orders. Both assignments were crossed, so it is possible to separately examine how each of them affects the breakoff and its timing. Chapter 4 will answer four research questions. The first two questions concern the breakoff timing and its impacting factors, and the last two focus on the filter questions and question topics.

1) When is the breakoff more likely to happen in the web survey?
2) What are the timing-varying predictors of the web survey breakoff?
3) Does the topic of the questions impact the breakoff and its timing?
4) Does the filter question format affect the timing of the survey breakoff?

2.4.2 During survey data collection

While most of the literature about breakoffs concentrate on factors impacting breakoffs, there is only one publication (Mittereder and West, 2021) researching the effects of intervening during the survey to minimise breakoffs. To develop an efficient real-time intervention system to prevent breakoff, three aspects must be well designed. These are: (1) the models for predicting the breakoff should be capable of generating accurate prediction at the question level, (2) the predictors included in the model should be coded in a way that maximises its predictive performance, and (3) the interventions should be effective in discouraging the breakoff. Given the scarcity of existing research in this area, all these three aspects need further investigation. Chapter 5 of this thesis focuses on the first two aspects: the choice of the predictive models and the coding of the predictors.

Currently, the most widely used model when studying breakoffs is the Cox survival model (Peytchev, 2009; Hochheimer *et al.*, 2016; Mittereder and West, 2021). This model estimates the probability of a respondent quitting the survey at a specific question given that the event has not happened yet (Singer and Willett, 2003). However, its proportionate hazard assumption, which states that the influence of a variable on the breakoff likelihood remains the same across different questions, is likely to be problematic, especially when this model is applied to the task of prediction. This is because the proportionate hazard assumption is often violated (see Mittereder and West, 2021 for an example), and using a model with a wrong assumption is unlikely to fit the data well and produce accurate prediction. Furthermore, the Cox survival model will use all the input variables as predictors even though some of them contribute little to the prediction. This might be problematic when multiple predictors, interaction terms, and non-linear effects are included in the model. As a result, the model is likely to suffer from the overfitting issue (i.e. the model fits the observed dataset too well to give a good prediction performance in future unseen datasets) (James *et al.*, 2013).

Machine learning models offer an alternative to the Cox survival model. Those models are non-parametric, meaning that there are few model assumptions (Buskirk *et al.*, 2018). Also, some machine learning models can automatically exclude variables that do not contribute to the prediction of the outcome (Signorino and Kirchner, 2018) and can implicitly take into account the interaction effect among predictors (Kern, Klausch and Kreuter, 2019). All of these characteristics mean that learning might be superior to the Cox survival model in the task of prediction.

However, it is unknown from the existing literature whether the machine learning models can lead to a superior prediction performance when being applied to the dataset with a clustered structure. To be more specific, because the breakoff is a question-level event, some of its predictors are time-varying, meaning that their values can change from question to question (e.g., the number of words in the question and the question response time). As a result, the breakoff data will have a clustered structure (questions are clustered within respondents). Many machine learning models assume a wide data setting (i.e. each row in the data represents an independent observation). Indeed, many studies that reported the superior performance of the machine learning models over the traditional logistic regression were conducted in the wide data setting (e.g., Buskirk *et al.*, 2018; Signorino and Kirchner, 2018; Liu, 2020). Currently, there is a limited amount of research that applies the machine learning models to predict question-level breakoffs in clustered data and compares their prediction performance with that of the Cox survival model.

Not only is there a need for better statistical models to predict breakoffs, the coding of the predictors also needs further research. This is especially true for those time-varying variables. While some researchers accumulated the value of the time-varying predictors from the beginning of the survey and used this coding in the model (Peytchev, 2009), others treated them concurrently (Vehovar and Cehovin, 2014). Each of these coding schemes is based on researchers' belief of how time-varying variables affect the breakoff. For instance, researchers adopting the accumulative coding assume that breakoffs happen due to the gradual accumulation of the response burden since the survey begins. On the other hand, the assumption behind the concurrent coding is that the breakoff event is more related to the response burden participants experience in the moment. Different coding schemes and the underlying assumptions are likely to affect the breakoff prediction performance, but few papers have compared the impact of different coding of time-varying variables.

Chapter 5 of this thesis will fill the above two research gaps by building seven predictive models along with four different ways of coding the predictors. By evaluating and comparing their prediction performance using multiple metrics, Chapter 5 will answer the following four questions:

1) Do survival machine learning models predict web survey breakoffs more accurately than the traditional Cox survival model?
2) What is the best classification model for predicting web breakoffs in clustered data?
3) Does the best performing survival model predict web survey breakoffs more accurately than the best performing classification model?
4) What is the best way to treat time-varying predictors of breakoffs in order to maximise the prediction performance?

2.4.3 After survey data collection

Apart from tackling the breakoff issue by design optimisation and real-time intervention, post-survey adjustment is another alternative. Given that survey breakoff can be viewed as a special case of survey nonresponse, methods to correct for the survey nonresponse bias can be applied in the context of breakoff bias adjustment. However, such an application received little research attention in the current literature, and there are two main gaps in this area.

To begin with, it is worth discussing whether breakoff should be corrected for in a step separate from the unit nonresponse adjustment. In practice, some survey organisations do not treat both survey outcomes differently (Bailey *et al.*, 2017; CRONOS team, 2018). However, after examining the relationship between breakoff and unit nonresponse, Peytchev (2011) concluded that both outcomes shared some impacting factors but breakoff had its own impacting factors as well. Such a finding implies that a separate nonresponse model should be built to compensate for breakoffs. The misalignment between the theory and practice warrants further research, which is surprisingly lacking in the literature.

The second existing gap refers to the statistical methods used in the breakoff compensation. As reviewed earlier, existing literature borrowed the idea of weighting from survey nonresponse bias correction to tackle the breakoff bias. Nevertheless, weights are a unit-level statistic whereas breakoff is a question-level event. It is unclear if weighting is effective in this context. Meanwhile, multiple imputation operates at the question level but has not been applied in the context of breakoff compensation. It is unknown whether the imputation method has a superior performance compared to the weighting approach in terms of reducing the breakoff bias.

To address both gaps, Chapter 6 of this thesis will develop a simulation. Different rates of breakoff and causes are simulated, after which four methods of dealing with the survey breakoff are applied. Two of the methods do not compensate for breakoff specifically while the other two do. All four methods will be applied to the simulated breakoff data to estimate the statistics of interest in the study. By comparing their deviation from the benchmark value across different breakoff rates and causes, Chapter 6 will answer the following two questions:

1) Does compensating for breakoffs separately in the post-collection adjustment help reduce the bias in survey estimates?
2) How is the effectiveness of the different breakoff compensation methods affected by different breakoff rates and mechanisms?

Chapter 3 Data and Methods

3.1 Data

3.1.1 Lightspeed web survey

Two surveys are analysed in this thesis. The first one is a cross-sectional, non-probability, web survey administered in the Lightspeed opt-in panel. This panel is managed by Kantar (a global commercial market research company) in the United States. According to the company, the recruitment of the panellists was conducted via traditional advertising as well as both internal and external affiliate networks. Once joining the panel, members can see different surveys distributed in the platform and decide what surveys to answer. Upon completing the survey, the respondents will receive reward points which can be accrued and redeemed later.

The web survey analysed in this thesis is secondary data and has two waves. The first wave was collected between September and October 2019 while the second one was carried out in October 2020. Both waves covered the same topics (described later in this section), but the first wave had more questions than the second one (196 vs. 126 questions). However, the respondents spent, on average, a similar amount of time (11 minutes) in both waves, mainly because there were many filter questions in the first wave and many of them did not apply to most respondents.

Both waves of the data were collected for a research topic that is irrelevant to the research questions in this thesis (For details about the study, see Eckman, 2021). In brief, the initial aim of the web survey was to investigate how an alternative design for the filter questions in the Consumer Expenditure Survey can impact respondents' answers. Consumer Expenditure Survey (CE) is conducted by U.S. Census Bureau every month to estimate U.S. consumers' expenditures and income. It relies on many filter questions (to cover many items that consumers might purchase in reality), but only the interleafed format is currently being used to present those questions. Meanwhile, it is difficult to set up an experiment in CE to estimate how respondents would have answered the filter questions about their purchases if the grouped format was used. Against this backdrop, the Lightspeed web survey was conducted to mimic the CE while embedding experiments about different question designs (described later in this section).

The web survey had questions about six of the many topics covered in CE: (1) respondents' demographic information, (2) characteristics of their housing units, (3) household income, (4) clothing purchase, (5) utilities payment, and (6) non-health insurance (e.g., vehicle and home insurance). Because the Lightspeed web survey analysed in this thesis is an opt-in survey (a type of non-probability surveys), the number of individuals invited to the survey is unknown and it is impossible to calculate the response rate. Additionally, the opt-in nature of the survey means that the respondent profile is likely to deviate from the general population. Indeed, the survey is dominated by white respondents (74%) and females (66%). According to the U.S. Census Bureau (2022), 59% of population are white and 50% are female in 2022. The biases in the Lightspeed sample might make the findings less generalisable.

However, this web survey is still considered suitable for the present research for two main reasons. Firstly, it has a substantial number of breakoff cases. Out of 3128 and 2370 respondents in the first and the second waves, 520 and 403 broke off, respectively. This leads to a breakoff rate of approximately 17%. This amount of breakoff helps the development of robust models for explaining and predicting breakoff in this thesis.

Another reason why this web survey is chosen is that the experiments embedded in the survey have not been analysed in terms of its impact on breakoff. This is the research gap the present thesis will bridge. In total, there were three experiments in the web survey. Stephanie Eckman (one of the co-authors of Chapter 4) designed all three experiments and commissioned Kantar to implement them in the Lightspeed Panel. Other two co-authors and I did not participate in the design of the experiments and were not aware of this dataset until its fieldwork was complete. After this, we applied for the access to the data, which was then granted by Stephanie.

Two of the three experiments were implemented in both waves: the first one was concerned with the order of question topics while the other was related to the format of filter questions. The third experiment was only carried out in the second wave and focused on manipulating the order of questions within the same topic.

The first experiment is related to the order of the question topics. As mentioned earlier, the web survey had six topics. Questions of the same topic were organised in the same block, resulting in six question blocks. Respondents went through the questionnaire from the first block to the sixth block. The topics of the first, second, and sixth question blocks always remained the same (they are: respondents' demographic information, the characteristics of their housing unit, and household income, respectively). In contrast, the topics in the third, fourth, and fifth question blocks were randomised between respondents' clothing purchase, utility payment and non-health insurance. This led to six possible orders of the question blocks (See Table 3.1). Participants were randomly allocated to one of the six orders upon reaching the first randomised question block (i.e. Block 3). There were 317 respondents (10% of the sample) who quit the survey prior to Block 3, so each of the six orders had approximately 15% of the sample.

Table 3.1. All possible orders of the question blocks in the Lightspeed web survey.

Group	**Block 1**	**Block 2**	**Block 3** [ab]	**Block 4** [ab]	**Block 5** [ab]	**Block 6**
1	Demographics	Housing	Clothing	Utilities	Insurance	Income
2	Demographics	Housing	Clothing	Insurance	Utilities	Income
3	Demographics	Housing	Utilities	Clothing	Insurance	Income
4	Demographics	Housing	Utilities	Insurance	Clothing	Income
5	Demographics	Housing	Insurance	Clothing	Utilities	Income
6	Demographics	Housing	Insurance	Utilities	Clothing	Income
7	Demographics	Housing	Unknown	Unknown	Unknown	Income

[a] The experiment about the format of filter questions was implemented in this block in both waves of the survey.

[b] The experiment about the order of questions was implemented in this block. Additionally, this experiment was only present in the second wave of the survey.

The second experiment was about the filter question format and implemented only in the three randomised question blocks (i.e. Block 3, 4 and 5). Respondents were randomly assigned to either the grouped (49% of the sample) or interleafed format (51% of the sample). Depending on the question block, there were five to six filter questions, each of which could lead to five follow-ups.

The third experiment manipulated the order of questions within the three randomised question blocks (Block 3, 4 and 5), but it was only present in the second wave. To be specific, the questions in the three randomised question blocks were ordered in one of the two ways: (1) high-frequency to low-frequency and (2) low-frequency to high-frequency. The frequency is determined by how often the respondents answered “yes” to the filter questions in these three question blocks during the first survey wave. The order of the questions within Block 1, 2 and 6 remained fixed. Again, respondents were randomly assigned to one of the two frequency groups. In the end, 49% of sample were allocated to the high-low frequency group while the remaining participants were in the low-high frequency group.

All experiments were crossed, so respondents could only be allocated to one of the 12 groups in the first wave (2 filter question formats × 6 block orders) and one of the 24 groups in the second wave (2 filter question formats × 6 block orders × 2 question orders). As the respondents could only see one of the designs, they were not aware of the experimental manipulation. Stephanie Eckman, the designer of the experiments and one of the co-authors of Chapter 4, has obtained the approval from the Institutional Review Boards of her research institute, which makes sure that study is designed and conducted in a way that protects the rights, welfare, and privacy of the participants.

The first wave of the web survey will be analysed in Chapter 4 of this thesis to answer the research questions about how filter question formats and question topics impact breakoffs. The second wave is not analysed in Chapter 4 because the data of this wave were not available yet during the writing of this chapter.

However, both waves of the web survey will be combined in Chapter 5 to develop models (including machine learning models) for predicting breakoffs. Combining surveys of the same topic and structure allows us to maximise the number of breakoff events in the data. This is especially helpful for the development of machine learning models because fitting those models will require the entire data to be split into multiple subsets and maximising the available breakoff events ensures that all subsets have a sufficient number of breakoffs.

However, there is an issue when combining two waves of data in Chapter 5. That is, there will be repeated measures for some observations in the combined data. This is because some Lightspeed panellists might answer both survey waves. Having repeated measures of the same observation violates one of the common assumptions in many statistical models (i.e. independence among observations) and can potentially damage the model fit. It is unknown how many panellists participated in both waves. However, some models fitted in Chapter 5 (e.g., survival model) are specialised in handling such a clustered data structure. Even though some models in Chapter 5 do not have this feature (e.g., gradient boosting), one of the research questions in Chapter 5 is focused on testing how those models perform in the clustered data (where questions are clustered within respondents). Therefore, the issue of repeated measures should not be of great concern in this thesis.

3.1.2 CROss-National Online Survey

The second survey analysed in this thesis comes from the sixth wave of the CROss-National Online Survey (CRONOS) panel. Like the Lightspeed panel, CRONOS is also an online panel but probability based. It was created to test the feasibility and efficiency of conducting a survey that is online, probability-based and cross-national (CRONOS team, 2018). Given that it was a pilot, this panel was set up only in three countries: Estonia, Slovenia and Great Britain. The target population of CRONOS is individuals who live in private households in the three countries and are at least 18 years old.

The sample of CRONOS was recruited from the eighth wave of the European Social Survey (ESS), which is an ongoing, face-to-face, probability-based and cross-national survey in Europe. To be more specific, after completing the survey in the eighth wave of ESS in 2016, respondents in Estonia, Slovenia and Great Britain were invited to join the CRONOS. If the participants did not have access to the internet, a tablet and an internet connection were offered to them for the duration of the project. The participants were informed about different aspects of the study (e.g., purpose, organisations involved, research teams' contact details) and their rights (e.g., voluntariness, withdrawals). Their consent was obtained prior to the data collection (Villar and Sommer, 2017) .

Due to the difference in the availability of the sampling frame and the implementation of the survey fieldwork, different sampling frames and strategies were used in the three participating countries. For Estonia and Slovenia, the population registry was used as the sampling frame. Stratified sampling was applied to this frame, and sample members were selected from different strata. Because Great Britain did not have a population registry suitable for sampling purpose, the postal address file of households was used. To select individuals from households, a three-stage sampling strategy was adopted. At the first two stages, systematic sampling was used to select the households. Then, the interviewers used a

Kish selection grid at the final stage to randomly select an eligible person from the chosen household (Villar and Sommer, 2017).

The data collection of CRONOS panel took place between December 2016 and February 2018. Over this period, a 10-minute welcome survey and six 20-minute surveys were issued to the panellists at an interval of approximately two months. As a result, there are six waves in the collected data (in addition to the welcome survey). An unconditional incentive of £5/ €5 in the form of a voucher was offered to the panellists of CRONOS along with the survey invitation (Villar and Sommer, 2017).

Each wave has different topics, and the topics in the sixth wave (i.e. the survey wave used in this thesis) included but were not limited to attitudes towards income equality, society fairness and political efficacy. This wave was conducted between January and February in 2018. There were approximately 98 questions in this survey wave, and respondents spent, on average, 26 minutes prior to survey breakoff or completion. In the end, 1812 people across the three countries responded to the survey in the sixth wave, resulting in a response rate of approximately 80%. The breakoff rate is 6%, meaning that 110 of the respondents broke off (CRONOS team, 2018).

The sixth wave of the CRONOS will be used in Chapter 6 to simulate breakoffs. It is chosen for two reasons. To begin with, the sample was recruited from ESS, from which multiple background information about the sample can be obtained (e.g., their demographics and voting history). A rick set of background information will allow us to fine-tune the simulation. For instance, we can choose how many and what variables are influential on the survey nonresponse but not breakoff (or vice versa). Secondly, by basing the simulation on a real-world dataset (as opposed to creating the simulation using pre-specified distributions), the resultant simulated data will mimic the data collected in the fieldwork to a large extent, which helps the ultimate application.

3.2 Methods

Different statistical models are applied to the two surveys described above to answer the research questions in this thesis. All of them can be classified into two classes of models, namely traditional statistical models and machine learning models. The former in this thesis includes the traditional Cox survival model and logistic regression while the machine learning models fitted in this thesis are LASSO Cox, LASSO logistic regression, Support vector machine, Random forest, and Gradient boosting.

Both classes of models have pros and cons. For the traditional statistical models, they are easy to understand. The coefficients routinely generated by the models (e.g., odds ratios in logistic regression or hazard ratios in the Cox survival model) make it easy to interpret the impact of the explanatory variables on the outcome of interest. Additionally, whether the association between the outcome and explanatory variables is genuine (as opposed to random chance) can be easily answered by the statistical significance tests and the resultant *p*-values. All of this facilitates the interpretability.

However, the advantages of traditional statistical models come at the expense of prediction performance. The traditional statistical models are usually based on some assumptions, such as the linear relationship between the logit transformed outcome and the covariates in logistic regression (Stoltzfus, 2011) and proportionate hazards in the traditional Cox survival model (Mills, 2011). They are not always met in reality (see Mittereder and West, 2021 for an example). Building models on assumptions that do not align with the data generating process will likely damage the goodness-of-fit of the model and prediction performance. Another related weakness is that the traditional statistical models cannot automatically handle the complex relationship between variables (e.g., non-linear effect, interaction). This issue can be solved by explicitly including the relevant terms in the model (polynomials, interaction, etc.), but this will require some expert knowledge on the research topic, which not all users have.

Unlike the traditional statistical models, machine learning models are non-parametric, so they make nearly no assumption about the relationship between the outcome and explanatory variables. Instead, they focus on learning the patterns in the data, including but not limited to the main effect, interaction, and non-linear effect. This flexibility therefore maximises the chance that the estimated model captures the true underlying outcome-predictor relationship, ultimately giving the machine learning model a good prediction performance. Another advantage of the machine learning model is that some of them (e.g., LASSO Cox and LASSO logistic regression) have a built-in feature to exclude variables that contribute little to the prediction from the model (Signorino and Kirchner, 2018). This feature helps the model parsimony, and more importantly, reduces the risk of overfitting (i.e. the model fits the observed dataset too well to produce a good prediction in future unseen datasets) (James *et al.*, 2013).

The machine learning model also has some limitations. Firstly, being non-parametric means that most machine learning models do not generate model coefficients. It is therefore difficult to interpret how the explanatory variables relate to the outcome. Secondly, the data-driven nature of the algorithm means that the machine learning models can easily overfit the data. Solving this issue will require a combination of multiple techniques, such as training/testing data split, cross-validation, and hyperparameter tuning (details about these techniques are discussed in the Method section of Chapter 5). For instance, the data-driven nature of the machine learning models means that it needs some hyperparameters to control the extent to which the data impact the model fit. These hyperparameters need to be tuned such that the final model not only fits the current data to a satisfactory degree but also predicts the future unseen data well. To identify such hyperparameter values, multiple candidate values must be trialled. Applying all these techniques together makes the model building process complicated.

As reviewed above, there are benefits and drawbacks in the traditional statistical models and the machine learning models. Therefore, both classes of models are fitted in this thesis to investigate which can give better breakoff prediction performance. The remainder of this section will give an in-depth overview of each model used in this thesis.

3.2.1 Traditional Cox survival model

The traditional Cox survival model is specialised in explaining whether (and if so when) the event of interest happens and what are its impacting factors (Singer and Willett, 2003). To build the Cox survival model, three elements need to be specified: the event of interest, time measurement and the starting point of the time. The binary survey breakoff (1 = survey breakoff; otherwise, 0) is the event of interest. Time is measured by the cumulative number of questions respondents have seen and treated as a discrete variable, which is an approach widely adopted in the past literature on survey breakoff (Peytchev, 2009; Mittereder and West, 2021). As every respondent starts the survey with the same question, the beginning of the time is the same for everyone (i.e. the first question in the survey).

According to Willett and Singer (1993), when the time in the survival data is measured as a discrete unit, the likelihood function for estimating the Cox survival model and the standard logistic regression is algebraically equivalent, so both models should generate the same coefficients. Also, they take the same model form shown below (so the logistic regression is not separately reviewed for brevity).

$$\ln\left(\frac{P_{iq}}{1-P_{iq}}\right) = \alpha_q + \beta_1 \boldsymbol{X}_{i1} + \beta_2 \boldsymbol{X}_{i2}(q)$$

The dependent variable in the traditional Cox survival model is the logit of the breakoff hazard. Essentially, it is a conditional probability of person i breaking off at question q given that this person has not broken off at any question prior to q (Singer and Willett, 2003). The logit breakoff hazard is explained by the baseline hazard α_q as well as the user-supplied covariates X. The former is a constant and represents the breakoff hazard at question q when all covariates in the model are zero or at the reference level. The coefficients associated with the user-supplied covariates X quantify the impact of those covariates on the logit breakoff hazard. Both the time-constant covariates $\boldsymbol{X}_{i1}$ (e.g., ethnicity) and the time-varying counterparts $\boldsymbol{X}_{i2}(q)$ (e.g., question word count) can be included in the model.

Once fitting the model, the hazard ratio will be used to interpret the model. It is the ratio between two hazards and obtained by exponentiating the model coefficients β. It quantifies the change in the breakoff hazard per unit difference in a specific covariate while controlling for others. A greater-than-one (less-than-one) hazard ratio means that the covariate is associated with an increased (decreased) chance of the breakoff occurrence. When the hazard ratio is equal to 1, there is no association between the covariate and the breakoff hazard (Mills, 2011).

The traditional Cox survival model is chosen for two reasons. To begin with, it can handle two unique features commonly seen in the breakoff data. One of them is the clustered data structure (i.e. questions clustered by respondents). Another feature is censoring, which takes place when the observations have not experienced the event of interest when the data collection ends (Schober and Vetter, 2018). As mentioned earlier, approximately 17% and 6% of the respondents broke off in the Lightspeed and CRONOS surveys, respectively. It means that the majority of respondents did not break off at all, making them censored cases. Without the breakoff timing, these cases will be treated as if they have missing event timing and excluded by most statistical models, so the sample size for the model development will be drastically reduced. The traditional Cox survival model, nonetheless, uses every case until the point where it is censored, thereby maximising the information in the model fitting process (Singer and Willett, 2003).

The second reason for using the traditional Cox survival model is that it not only considers the time but also allows the effect of covariates on the breakoff to vary across time. This is a good fit to the main research question in Chapter 4 (i.e. how filter question formats and question topics impact breakoff and its timing).

3.2.2 LASSO Cox

LASSO Cox is a machine learning model fitted in this study. The major difference between this model and the traditional Cox survival model is that this model uses a penalty term during the fitting process. This term is referred to as λ, and it penalises the model that has many covariates (Tibshirani, 1997). The penalty term λ is a non-negative hyperparameter, and a larger λ will lead to more penalisation, which further results in some model coefficients shrinking towards zero. When the λ is large enough, some coefficients will become zero, and covariates with the zero coefficient will be automatically excluded from the model. As a result, the model becomes simpler. On the contrary, when λ is zero (i.e. its minimum value), no penalisation is applied, and the fitted LASSO Cox model is the same as the traditional Cox survival model. In the same vein, the LASSO logistic regression differs from the standard logistic regression due to the penalty term, so it is not reviewed here for brevity.

The LASSO Cox is fitted in this study for two reasons. The traditional Cox survival model is known to include all covariates in the model even though some of them might contribute little to the prediction of breakoffs. This will damage the interpretability of the result and increase the risk of overfitting. LASSO Cox is used mainly to investigate whether the simpler model can predict the breakoff more accurately than the full-size Cox survival model (one of the research questions in Chapter 5). Another reason for developing the LASSO Cox is that there are a large number of predictors available in the data, and the penalisation feature of this model can shed light on what predictors are more predictive of the breakoffs.

3.2.3 Support vector machine

Support vector machine (SVM) is a machine learning model. This model splits the breakoff cases from the complete respondents using a hyperplane in the high-dimensional space defined by the number of predictors (Kirchner and Signorino, 2018). Given that breakoff cases and complete respondents usually cannot be separated linearly or perfectly, two solutions are used in SVM, and they are the two main hyperparameters to tune in SVM (Rhys, 2020).

The first solution is to apply some transformations to the predictors. As such, the space in which SVM operates is enlarged, and a linear separation between breakoff cases and complete respondents becomes possible in this new space. The function for the transformation is called the kernel $\varphi(\cdot)$, and it is one of the hyperparameters in SVM. There are three commonly used kernels: linear kernel (i.e. no transformation), polynomial kernel (e.g., quadratic, cubed which is controlled by the degree of the polynomial term *d*), and radial kernel (which has its own hyperparameter σ to control the influence of each observation on the position of the hyperplane). Increasing *d* and σ will make the model more flexible and fit the observed data better, but the risk of overfitting rises as well. Therefore, they need to be tuned such that the model not only fits the present data reasonably well but also gives a good prediction performance in the future unseen data.

Another solution to the problem of imperfect separation in SVM is to allow misclassification. That is, some observations are allowed to be incorrectly predicted by SVM as (non-)breakoff cases. Th hyperparameter C will control the extent to which the misclassification is allowed. It is a non-negative term, and a larger value will impose more penalty on the misclassified cases. As a result, the model algorithm will focus more on making the prediction of those cases correct. However, if the C value is too large, the risk of overfitting will increase. Therefore, researchers have to tune this hyperparameter to achieve a good balance between the amount of allowed misclassification and correct prediction.

3.2.4 Random forest

Random forest is another machine learning model fitted in this study. It requires the development of multiple decision trees, each of which recursively splits the respondents into two child nodes using one of the input predictors (Lantz, 2019). As the aim of the split is to make the cases within the same child node become less heterogeneous (or more homogeneous), the predictor that leads to the largest reduction in the heterogeneity from the parent node will be chosen by the algorithm (Buskirk, 2018). The heterogeneity is measured by the Gini index. This index is calculated based on the proportion of breakoff and non-breakoff cases in the node, and a smaller value is preferred as it means that the node contains many cases of the same breakoff status (James *et al.*, 2013).

There are three main hyperparameters to tune in random forest (Rhys, 2020). The first hyperparameter is the minimum number of cases in a node for a split to continue (commonly denoted as *min_n*). It is a positive number. The smaller this number is the higher the risk of overfitting will be. This is because the decision tree will focus too much on some specific individual cases. Meanwhile, if *min_n* is set too high, the goodness-of-fit of model will be adversely affected as the tree is prevented from conducting further necessary splits to make the cases in the node more homogenous.

The number of trees in the random forest is the second tuning hyperparameter (denoted as *trees*). When fitting the random forest, B bootstrapped samples will first be randomly drawn from the original data. This procedure is to add variations to the data and reduce the risk of overfitting. Each bootstrapped dataset will have as many observations as the original data, and the decision tree will be developed independently in those B bootstrapped datasets, leading to B trees in the forest. Once the random forest is developed, each tree will generate its own prediction for respondent's breakoff status, and the most frequently predicted status will be the final prediction. If too few bootstrapped samples are drawn, a small number of trees will be in the forest. As a consequence, there will be large variations in the predicted outcome, meaning that the prediction becomes less consistent and reliable. In contrast, if a large value is chosen for the trees, the model development will be less efficient because of the diminishing return (i.e. fitting more trees will consume more time but the resultant gain in the prediction performance is diminishing).

The third hyperparameter in random forest is the number of predictors to consider when conducting a split (denoted as *mtry*). Considering all predictors in the split will likely see some specific predictors always being chosen to make the split. This makes the trees similar to each other and the final random forest model less robust to different datasets. To diversify the trees in the forest, only a random subset of the input predictors will be considered at each split, thereby forcing the tree to be different. The value of this hyperparameter ranges from one to the number of input predictors. A large value for *mtry* will make the trees similar and lead to the issue discussed earlier. On the other hand, when a low value is used for *mtry* and some input predictors are not associated with the outcome variable, the tree might have to use those uninformative predictors, meaning that the model will have some splits which do not contribute to the prediction.

3.2.5 Gradient boosting

Like the random forest, gradient boosting is also based on multiple decision trees. The main difference between them is how the decision trees are developed. While the decision trees in random forest are fitted independently from each other, gradient boosting develops the trees in a sequential and iterative manner (Hastie, Tibshirani and Friedman, 2009). Therefore, the trees in previous iterations affect the subsequent one. The overall idea is to build a weak decision tree model (which outperforms the random chance only slightly) to predict the breakoff at each iteration and then develop trees in the subsequent iterations to gradually correct the prediction errors made by the previous trees. In the end, although the model in each iteration is weak, combining them together will lead to a strong predictive model (Mayr *et al.*, 2014). The sum of the predictions made by each tree will be the predicted breakoff hazard for the respondent.

To develop the gradient boosting, the algorithm begins by assigning each respondent the same constant (e.g., the average breakoff hazard from the collected sample). It then subtracts this constant from each respondent's true breakoff status (i.e. 0/1) to obtain the prediction errors. These errors are used as the dependent variable in the first decision tree. After this, the first decision tree will generate its prediction, which is then added together with the initially assigned constant to form a new set of predicted breakoff hazard. The difference between the new prediction and the true breakoff status will update the prediction errors. The algorithm will proceed to fit the second decision tree using the updated prediction errors. This process (i.e. fitting new decision trees on prediction errors, combining the initially assigned constant and predictions from all the trees fitted so far, making new predictions, updating prediction errors) will continue until some pre-specified conditions are reached. Those conditions are controlled by the hyperparameters in gradient boosting.

Both gradient boosting and random forest are the tree-based model, so they have some hyperparameters in common. They are the minimum number of cases in a node for a split to continue (*min_n*), the number of trees in the model (*trees*), and the number of predictors to consider when making a split (*mtry*).

In addition to the three hyperparameters above, the gradient boosting has two special hyperparameters. The first one is how many splits a tree can have (*tree_depth*). Unlike random forest where each decision tree is fully developed, trees in gradient boosting are only developed to a certain depth. This is because a fully developed tree (i.e. a large value in *tree_depth*) can easily lead to the overfitting issue, and the dependency between trees in gradient boosting means that those individual overfitted trees will have an adverse impact on the final model. On the other hand, a small value in *tree_depth* can be computationally demanding as it will require many decision trees to be fitted (i.e. a large value in the *trees* hyperparameter). Also, if not enough number of trees is specified, the final model might become suboptimal as it underfits the data.

The second hyperparameter in gradient boosting is called the learning rate (*learn_rate*), which controls how quickly the subsequent tree learns from the prediction errors made by the previous trees. Its value ranges from zero to one. A small value is usually preferred because the algorithm that learns slowly from the errors tends to perform better when predicting unseen data (Natekin and Knoll, 2013). However, a smaller value will require more trees and therefore more computation time. When the number of trees to be fitted is set too low, a slow learning rate might also lead to a suboptimal model.

Logistic regression, LASSO logistic regression, Support vector machine, Random forest and Gradient boosting are five models that cannot handle the clustering structure in the breakoff data. The logistic regression is traditionally used in the study of survey breakoff (e.g., Tijdens, 2014; Blumenberg *et al.*, 2018) and can be used as a benchmark for the other four models. The other four models are chosen because they were found to have a superior performance in predicting survey nonresponse over the logistic regression in many existing studies (e.g., Buskirk, 2018; Kirchner and Signorino, 2018; Signorino and Kirchner, 2018; Liu, 2020; Kern, Weiß and Kolb, 2021). Meanwhile, given the lack of study on what models are more predictive of the imminent breakoffs during the survey process, it is necessary to test multiple models. Indeed, applying these five models to the breakoff data fits into one the focuses of Chapter 5 (i.e. understanding how those models perform when the clustering of the data is ignored and identifying what model among them is most predictive of the breakoff). Moreover, the best performing machine learning model can be compared to the best performing survival model to answer another research question in Chapter 5 (i.e. whether or not taking into account the special data structure in the breakoff data by the survival model helps improve the breakoff prediction).

Chapter 4 Impact of Question Topics and Filter Question Formats on Web-survey Breakoffs

Abstract

Web surveys have become increasingly popular over the last decade, but they tend to suffer from breakoffs, which take place when respondents start the survey but do not complete it. Many studies have investigated the factors impacting breakoffs, but they often ignored the breakoff timing and gave scant attention to two factors: question topics and filter question formats (grouped vs. interleafed as defined by whether filter questions are presented upfront or not). Using survival analysis, this study first identifies when breakoffs are more likely to happen and what are the time-varying predictors of breakoffs. Then, by using a web survey that experimentally manipulates the filter question format and randomly orders the question topic, this study investigates the effect of question topics and filter question formats on the breakoff event and its timing. We find that most breakoffs tend to happen at the beginning of the questionnaire and at the place where a new question topic is introduced. While item nonresponse is associated with more breakoffs, it is surprising to see that open-ended and long questions are associated with a lower breakoff risk. Additionally, we discover that grouping the filter questions leads to fewer breakoffs at the beginning compared to the interleafed counterpart, but the breakoff risk in the grouped format catches up quickly when respondents realise their previous answers will trigger more questions. This study also shows that questions about insurance have more breakoffs while questions on demographics and income have fewer breakoffs despite their sensitivity level.

Keywords

Web surveys, breakoff timing, survival model, questionnaire design, experimental design

4.1 Introduction

Surveys have been widely used in different fields, such as market research and political polling. Due to the cost concern and tight schedule, an increasingly number of surveys have been conducted online. However, running surveys on the web has some limitations, one of which is survey breakoff. Survey breakoffs happen when the respondent starts the survey but fails to complete it (Tourangeau, Conrad and Couper, 2013). As a result, missing data are produced, causing subsequent analysis to have lower statistical power as well as potentially biasing results (Steinbrecher, Roßmann and Blumenstiel, 2015).

To address the breakoff, it is important to understand its determinants. Past studies (see Peytchev, 2009 for an example) have identified many factors impacting breakoffs, but most of them studied breakoffs only as a binary outcome and ignored the breakoff timing. Investigating breakoff timing is important as survey practitioners want respondents not only to complete the survey but also to complete as many questions as possible before they break off (Sakshaug and Crawford, 2010).

The present study will apply survival analysis to an opt-in web survey to investigate when breakoffs are more likely to happen and what are the time-varying factors (factors whose value varies throughout the questionnaire) that explain breakoffs. Additionally, the study will investigate two other important factors that have received scant attention: question topic and the format of filter questions.

The content, sensitivity and placement of question topics can impact the breakoff and its timing. Topics that are relevant to the respondents can decrease or postpone breakoffs (Shropshire, Hawdon and Witte, 2009) while sensitive topics might have the opposite effect. When studying the effect of question topics on breakoff, randomising the topic order is important; otherwise, ignoring the ordering effect could confound the topic effect and cause spurious correlations with breakoffs. Nevertheless, this has not been done in prior research.

The format of filter questions is another factor that has received limited attention in the breakoff literature. Filter questions can produce a high degree of response burden as the positive answer to a filter question can lead to more questions. There are two main ways of presenting filter questions and their follow-ups. In the **grouped format**, all filter questions are asked before the follow-ups are displayed whereas in the **interleafed format** every filter question immediately triggers its follow-ups (Kreuter, Eckman and Tourangeau, 2020). Although both filter formats can cause response burden and are prone to breakoffs, there is a difference between them in the timing when respondents learn about the burden. Respondents answering the interleafed format will quickly understand the response burden after giving affirmative answers to one or two filter questions. They could break off as early as the first pair of filter and follow-up questions. In the grouped format, respondents can only learn about the extra burden when they reach the follow-ups. They are therefore expected to break off later. However, no previous research has tested the relationship between breakoff timing and filter question formats.

This study uses a web survey that experimentally manipulated the filter question format and randomly ordered the question topic. Thus, we are able to causally investigate the impact of these two factors on the breakoff and its timing.

4.2 Background

4.2.1 Framework for studying breakoffs

Breakoffs are prevalent in web surveys. For example, Revilla (2017) reviewed 185 opt-in web surveys distributed through a Spanish survey company and found that the mean breakoff rate was 11.8%. In an online probability survey about University of Michigan staff and students' attitudes towards environmental issues, the breakoff rate was 13%, 14% and 17% in the year of 2014, 2015 and 2018, respectively (Mittereder, 2019).

Given the prevalence of survey breakoff, many researchers have been studying its impacting factors. As a result, a framework has been developed to summarise different factors. According to Peytchev (2009) and Mittereder and West (2021), these factors can be grouped into four categories: (1) page/question characteristics, (2) survey design, (3) respondent factors and (4) paradata.

Page/question characteristics refer to the design features of survey pages and questions. Cognitively demanding questions such as matrix, open-ended questions and questions with more characters are associated with more breakoffs (Peytchev, 2009; Hoerger, 2010; Tijdens, 2014; Steinbrecher, Roßmann and Blumenstiel, 2015). These types of questions can impose extra burden on respondents when they engage in a series of actions required to answer the question such as comprehending the question and retrieving the relevant information (Tourangeau, 2018). To avoid the burden, respondents might choose to break off.

The second factor that is related to breakoff is the survey design. Examples of this are providing incentives unconditional on survey completion (Silber, Lischewski and Leibold, 2013), using a lengthy questionnaire (Hoerger, 2010) and displaying the progress bar alongside the questionnaire (Villar, Callegaro and Yang, 2013).

The third group of factors - respondent factors - refers to the characteristics of the respondents. They are used in the existing literature as proxies for sample members' tendency to cooperate with the request to survey response (Durrant and Steele, 2009) or their cognitive ability to handle the burden from the survey response (Roßmann, Gummer and Silber, 2018). Survey breakoff is conditional upon survey response, so both are similar in nature and likely to be correlated. Therefore, respondent factors that are associated with the survey response are often used to explain survey breakoff. For example, age and education are two respondent factors that are commonly used to represent individuals' ability to cope with the survey burden. In fact, respondents who are older and have a lower education degree were found to be more likely to break off (Peytchev, 2009; Blumenberg *et al.*, 2018). Another example is related to factors that reflect respondents' general level of cooperation with the survey request. The findings regarding the relationship between these factors and breakoffs are often mixed, including gender, race, marital status, student and income (Galesic, 2006; Peytchev, 2009, 2011; Klein *et al.*, 2011; Mittereder and West, 2021), but the general trend is that male, non-white, student and more affluent respondents are more likely to break off.

Paradata, the final category in the framework, refer to the information collected during the response process (Kreuter, 2013). This type of data is believed to reflect the change in the response burden and respondents' motivation throughout the questionnaire (Mittereder and West, 2021), thereby being useful for predicting the imminent breakoff. Some paradata that have been associated with breakoffs are the proportion of questions that are not answered (Mittereder and West, 2021) and using mobile devices to answer the survey (Wenz, 2017).

The above four categories form a comprehensive framework. However, previous research has found that web survey breakoffs were preceded by an accumulated respondent burden

(Galesic, 2006). This means that the burden caused by the factors in the framework takes some time before it can actually exert its influence on breakoffs. An example of this can be seen in the study conducted by Mittereder and West (2021). By allowing the effects of the responding device to vary in time (measured as the cumulative number of questions answered), they found that there was no difference in the breakoff between the non-mobile and mobile devices at the beginning of the survey. However, when mobile device respondents answered more questions, they were more likely to break off.

Based on the review so far, we argue that the timing dimension is a necessary factor to consider in the study of breakoffs and the effect of time-varying factors on the breakoff needs further research. In the present study, we derive the time-varying factors from question characteristics (e.g., question word count) and paradata (e.g., item nonresponse rate) and investigate how they impact breakoffs, after controlling for the difference in respondents' cognitive ability and survey cooperation using their demographic information such as gender, age, ethnicity and education. The first two research questions are:

RQ1. When is the breakoff more likely to happen in the web survey?
RQ2. What are the timing-varying predictors of the web survey breakoff?

In addition to examining breakoff timing, this study will contribute to the literature by focusing on two specific factors: question topic and the filter question format.

4.2.2 Question topics and breakoffs

Many studies have identified the survey topic as an important factor for unit nonresponse (not answering the survey at all) (Groves, Singer and Corning, 2000) and item nonresponse (not answering some of the questions) (Tourangeau and Yan, 2007). Survey breakoff, as a special type of nonresponse, is also impacted by respondents' topic interest. For instance, when analysing the data from a web survey that covered the topic of conservation, Shropshire, Hawdon and Witte (2009) documented that respondents who scored higher in their conservation support were less likely to break off.

In addition to the interest in the topic, the perceived sensitivity of the topic can also impact survey nonresponse. When facing a sensitive topic such as income or sexual orientation, respondents might feel uncomfortable with its intrusiveness or worried about the potential threat of disclosing personal information (Tourangeau and Yan, 2007). As a result, respondents will skip those sensitive questions, leading to item nonresponse. In fact, the behaviour of skipping sensitive questions was found to be more frequent in the interviewer-administered survey mode compared to a self-completion mode (Kreuter, Presser and Tourangeau, 2008).

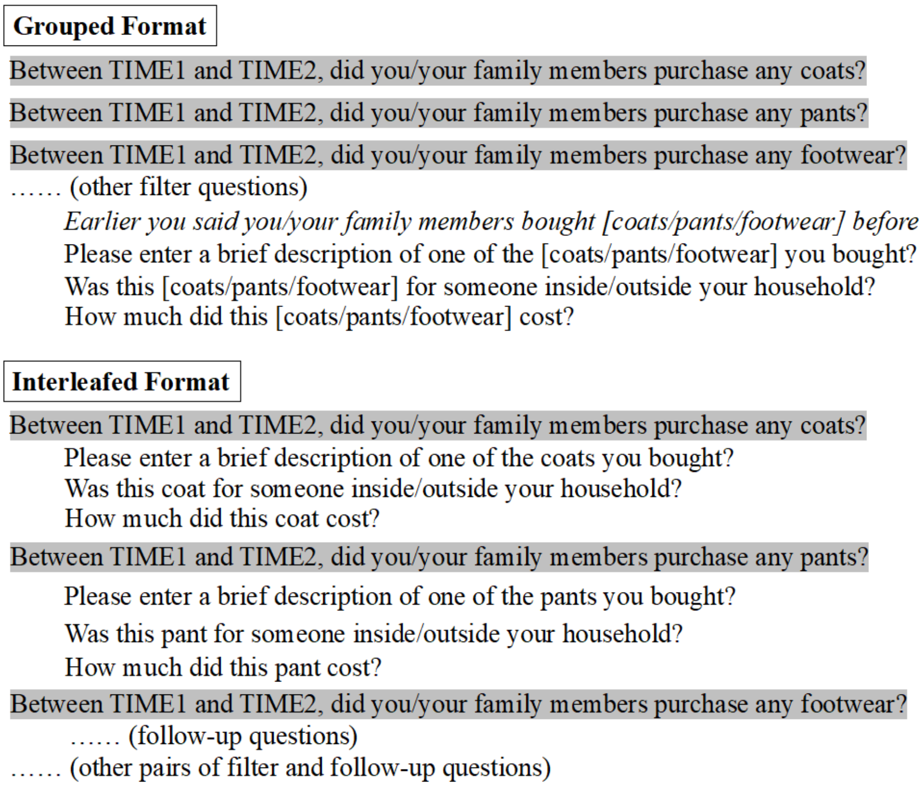
As an alternative to not answering the sensitive question, respondents might terminate their survey participation. However, among those papers that investigated the relationship between question topics and breakoffs (see McGonagle, 2013; Mittereder and West, 2021), the order of the topics was not randomised. For example, McGonagle (2013) analysed a telephone survey about the U.S. families' economic status, but the topic about respondents' housing was always followed by their employment history, income and so on. Previous research has noted that the order of the topic could affect the rates of item nonresponse (Teclaw, Price and Osatuke, 2012).

Without the order randomisation, prior work failed to separate the impact of the topic content from that of the topic order. An ideal design to investigate the impact of question topic on breakoff is to randomise the order of questions and include questions with different levels of sensitivity. We used such a design to answer our third research question:

RQ3. Does the topic of the questions impact the breakoff and its timing?

4.2.3 Filter question formats and breakoffs

The use of filter questions can also impact web survey breakoff and its timing. Many surveys use filter questions which trigger some follow-up questions when answered positively. For example, if the respondent chooses "yes" to the filter question "Have you held a full-time job during the past 12 months", then more questions will follow (e.g., "From when and until when did you hold this job"). The grouped and interleafed formats are two main ways to present filter and follow-up questions. A visual example of both formats is shown in Figure 4.1.



Grouped Format

Between TIME1 and TIME2, did you/your family members purchase any coats?

Between TIME1 and TIME2, did you/your family members purchase any pants?

Between TIME1 and TIME2, did you/your family members purchase any footwear?

…… (other filter questions)

Earlier you said you/your family members bought [coats/pants/footwear] before
Please enter a brief description of one of the [coats/pants/footwear] you bought?
Was this [coats/pants/footwear] for someone inside/outside your household?
How much did this [coats/pants/footwear] cost?

Interleafed Format

Between TIME1 and TIME2, did you/your family members purchase any coats?

Please enter a brief description of one of the coats you bought?
Was this coat for someone inside/outside your household?
How much did this coat cost?

Between TIME1 and TIME2, did you/your family members purchase any pants?

Please enter a brief description of one of the pants you bought?
Was this pant for someone inside/outside your household?
How much did this pant cost?

Between TIME1 and TIME2, did you/your family members purchase any footwear?

…… (follow-up questions)

…… (other pairs of filter and follow-up questions)

Figure 4.1. Example of grouped and interleafed formats in the web survey analysed in this study (filter questions are highlighted).

One advantage of the grouped format is that the connection between a "yes" answer to the filter questions and the activation of follow-ups is not immediately apparent, so respondents facing the grouped filter questions would choose more "yes" answers compared to the interleafed format. This was found in Eckman *et al.* (2014) after randomly assigning the respondents of a probability-based telephone survey to either grouped or interleafed formats and comparing the number of "yes" in filter questions between the two formats.

However, in the grouped version, the follow-up questions are far away from the corresponding filter questions, so respondents have to recall the relevant information from their memory again, which causes recall difficulties and hampers the cognitive processing (Clark-Fobia, Kephart and Nelson, 2018). Kreuter *et al.* (2011) also randomly allocated sample members of a different telephone survey to the grouped or interleafed format and noted that respondents in the grouped format chose more non-substantive answers (e.g., "Don't know") for the follow-up questions.

Unlike the grouped format, the interleafed version puts together questions that are of the same topic, serving as a recall aid (Kreuter, Eckman and Tourangeau, 2020). Yet, in the interleafed version, respondents can quickly learn that a positive answer to the filter question will trigger more questions. They then are more likely to deliberately choose a "no" to shorten the questionnaire, which was documented in the Eckman *et al.* (2014) and Kreuter *et al.* (2011) studies mentioned above.

The review of the grouped and interleafed formats highlights that both formats impose burden on the response process and respondents who do not want to or cannot handle this burden will provide lower data quality. Rather than giving incorrect answers to reduce the length of the survey, respondents could break off. In addition to comparing the effect of grouped and interleafed formats on measurement error, Kreuter *et.al* (2011) and Eckman and Kreuter (2018) also looked at the influence of the filter question format on breakoffs. Both studies found that the format was not associated with breakoffs.

However, previous studies did not investigate whether there is a difference in the breakoff timing between filter question formats. As we argued previously, it is important to consider the timing in breakoff studies as it might produce new insights regarding mitigating breakoffs. Thus, the final research question in this study is:

RQ4. Does the filter question format affect the timing of the survey breakoff?

4.3 Data

The data used in this study come from a web survey conducted between September and October 2019. The web survey was administered to members of the Lightspeed Panel, an opt-in web panel in the United States. Upon completing the survey, the respondents received reward points which could be accrued and redeemed later. Given the opt-in nature, it is impossible to calculate the response rate. The survey analysed here is dominated by white respondents (74%) and females (66%). These deviations from the US population make the subsequent findings less generalisable.

Nevertheless, we consider that this web survey is appropriate to be analysed for three reasons. First, it records the outcome of interest - breakoffs. After removing two individuals with an unknown response status, the final sample size for analysis is 3,128. Out of these, 520 respondents accessed but did not complete the survey, resulting in a breakoff rate of approximately 17%. This breakoff rate is slightly higher than other surveys reviewed in the previous section even though individuals voluntarily participated in this survey and could only receive the reward upon survey completion. Meanwhile, the survey recorded the last question respondents saw, enabling the investigation of breakoff timing.

Secondly, the web survey includes six different topics with varying levels of sensitivity. Questions of the same topic are organised into a single block, resulting in six blocks in the survey. As shown in Table 4.1, the topic of Block 1, 2 and 6 always remains the same, namely respondents' demographic information, housing unit and household income. In contrast, the topics of the three remaining blocks (Block 3, 4 and 5) are randomised among respondents' clothing purchase, utilities payment and non-health insurance (e.g., vehicle and home insurance). This randomisation leads to six possible orders among the blocks (See Table 4.1). Respondents were randomly assigned to one of the six orders upon seeing the first randomised block (i.e. Block 3). For 317 respondents who broke off at Block 1 or 2, their assigned order is unknown. Within each block, the order of the questions is fixed.

Table 4.1. All possible orders of the question blocks in the web survey analysed in this study.

Order	**Block 1**	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**Block 6**	**Breakoff/ Total**
1	Demographics	Housing	Clothing	Utilities	Insurance	Income	21/461
2	Demographics	Housing	Clothing	Insurance	Utilities	Income	32/470
3	Demographics	Housing	Utilities	Clothing	Insurance	Income	27/476
4	Demographics	Housing	Utilities	Insurance	Clothing	Income	36/454
5	Demographics	Housing	Insurance	Clothing	Utilities	Income	48/478
6	Demographics	Housing	Insurance	Utilities	Clothing	Income	39/472
7	Demographics	Housing	Unknown	Unknown	Unknown	Income	317/317
Total							**520/3128**

Different topics in the survey help answer the research question regarding how question topics affect breakoffs, especially for the three topics whose order is randomised. Also, Demographics and Income (Block 1 and 6) are considered sensitive as they are either found to suffer from more item nonresponse (Tourangeau and Yan, 2007) or recommended by survey practitioners to be placed towards the end of the questionnaire so respondents feel more comfortable to share such information (Allen, 2017). This varying sensitivity allows us to study the relationship between topic sensitivity and the survey breakoff.

Lastly, the survey embedded an experiment about filter question formats in the three randomly ordered blocks (i.e. Block 3, 4 and 5). The respondents were randomly assigned to either the grouped (49% of the sample) or the interleafed format (51% of the sample). Depending on the block, there are five to six filter questions, each of which can trigger five follow-ups.

In total, the survey analysed here has 196 question items, and approximately 80% of them are in the three randomised blocks (See Table A.1 in Appendix A for the number of questions in each block). At the beginning of nearly every question block, there is an introduction statement which informs respondents of the upcoming block's topic and encourages respondents to give accurate answers. Respondents can either click a radio button to show their acknowledgement or skip to the next question. Among the 196 total items in the questionnaire, six items are introduction statements, which we code as the reference category for the question topic. On average, the respondent who broke off saw 16 questions (standard deviation = 19), much lower compared to those completed the survey (85 questions, standard deviation = 21). The descriptive summary for all variables used in this study along with how they are coded are provided in Table A.1 and Table A.2 of Appendix A.

4.4 Method

We use the survival model to answer the research questions. The survival model is useful in explaining whether, and if so when, the event of interest happens (Singer and Willett, 2003). Following previous studies on survey breakoffs (Peytchev, 2009; Mittereder and West, 2021), time is measured by the cumulative number of questions respondents saw and treated as discrete. As Willett and Singer (1993) emphasised, when the time metric is discrete, the likelihood function for estimating the discrete-time survival model and the standard logistic regression is algebraically equivalent. We therefore use the standard logistic regression to fit the discrete-time survival model in this study. The model is estimated using the *glm* command in R 4.0.2 (R Core Team, 2020) and takes the following form:

$$\ln\left(\frac{P_{iq}}{1-P_{iq}}\right) = \alpha_q + \beta_1\boldsymbol{X}_{i1} + \beta_2\boldsymbol{X}_{i2}(q)$$

P_{iq} represents the probability of person i breaking off at question q given that this person has not broken off at any question prior to q. This conditional probability is called hazard in the survival literature (Singer and Willett, 2003). The equation shows that the logit transformed hazard is a linear function of three terms. The first term, α_q, is the baseline hazard, which quantifies the hazard of breaking off at question q when all covariates in the model are zero. The other two terms, $\beta_1\boldsymbol{X}_{i1}$ and $\beta_2\boldsymbol{X}_{i2}(q)$ represent a set of different covariates X and their impact β on the logit hazard. The difference between them is that $\boldsymbol{X}_{i1}$ represents the time-constant covariates (e.g., ethnicity) and $\boldsymbol{X}_{i2}(q)$ represents the time-varying counterparts (e.g., question word count).

Four logistic models are developed to answer the research questions in this study.[1] Model 1 involves only time represented as the number of questions seen and the respondents' demographic characteristics while Model 2 adds in the time-varying factors. These two models will together address RQ 1 and 2 (i.e. when breakoffs are likely to happen and what are the time-varying predictors).

To answer RQ 3 and 4 (i.e. how question topics and filter question formats affect breakoff and its timing), the analysis sample is restricted to only Blocks 3, 4 and 5. As mentioned earlier, the experiment of topic orders and filter question formats only exists in these three blocks. The sample restriction enables us to only investigate the breakoffs happening under the experimental design and measure the effect of both factors on breakoffs more directly. Model 3 is derived by applying Model 2 to the restricted sample but with two changes. Firstly, given that some topics are discarded in the restricted sample, the variable about question topics now includes only four categories, namely Clothing (the reference category), Utilities, Insurance and Introduction Statement. Using Clothing as the reference category (rather than the Introduction Statement as in Model 1 and 2) helps investigate how the topics of other two randomised blocks (Utilities and Insurance) impact breakoffs in comparison to

[1] We also fitted the continuous-time survival model using R's *survival* package (Therneau, 2021), but both survival and logistic models gave the same result (not shown). We report the logistic regression in this study as it is the model used in many fields for discrete time events.

Clothing. Secondly, the variable representing the matrix questions is excluded as these questions only exist in Block 2, which is eliminated from the restricted sample. In Model 4 we add in two interaction terms between time (i.e. number of questions seen) and the grouped/interleafed format as well as the question topics. Model 3 will investigate whether the question topics and filter question formats impact the breakoff risk, and Model 4 will answer whether their impact on breakoffs changes throughout time.

Due to a sizable number of breakoffs prior to the demographic-related questions, demographic variables suffer from missing data (ranging from 2% to 8% as shown in Table A.1 in Appendix A). To fill in the breakoff cases' missing demographic information, we used multiple imputation. Following Enders (2010), we included all variables in the substantive model in the imputation (i.e. breakoff status, time, demographics, question characteristics and paradata) as well as the order of question blocks respondents were assigned to (as shown in Table 4.1). We created 10 imputed datasets, each of which was obtained after 50 iterations. Parameters of all substantive models were separately estimated on these 10 datasets and then pooled together by the combining rule of Rubin (1987).

By including the breakoff status in the imputation for the missing demographics, the association between them might be artificially increased. As a result, the influence of some demographic variables might be inflated in the substantive models that are developed later for explaining breakoffs. However, this approach was recommended in the literature on multiple imputation (Sterne *et al.*, 2009). Not including the breakoff status in the imputation of demographics will assume that there is no relationship between them. This assumption is wrong as many past studies have already documented that individuals with certain characteristics are more likely to break off, such as those who are older and have a lower education degree (Peytchev, 2009; Blumenberg *et al.*, 2018). Therefore, using the breakoff status during the imputation helps preserve the relationship between breakoff and demographics. Secondly, the present study is focused on investigating the impact of filter question formats and question topics on breakoff, so the potential bias in the model coefficients related to the demographics is tolerated.

To understand how different methods for handling the missing demographics can impact the model result, we also coded the missingness in demographics variables explicitly as a category in the model in a sensitivity analysis, but the conclusion regarding our research questions does not change (See these results in Table A.3 and Table A.4 of Appendix A). Therefore, models built upon the imputed datasets are reported here. The imputation was performed in R 4.0.2 using the *mice* package (van Buuren and Groothuis-Oudshoorn, 2011). For the univariate description of the variables before and after imputation see Table A.1 in Appendix A.

4.5 Results

4.5.1 Change in breakoff hazard over time

Figure 4.2 plots time (i.e. the number of questions seen) on the *x* axis and the breakoff hazard on the *y* axis. A larger hazard indicates a higher breakoff risk. Figure 4.2 illustrates that the largest breakoff hazard is at the beginning of the survey. The second peak lies between the 15^{th} and 20^{th} questions. Because questions in the range of the second peak either involve sensitive topics (i.e. rent/mortgage for the dwelling), belong to matrix questions or introduce a new series of topics, we speculate that the second peak is more likely attributed to the question characteristics rather than time. After the second peak, the breakoff hazard tapers off. All peaks after 100 questions are mainly due to the rare breakoff event and decreasing number of respondents included in the denominator for calculating the breakoff hazard (For instance, at the 115^{th} question, only 206 respondents remained in the survey, and a single breakoff event among this small denominator is causing large peaks in the tail of the distribution).

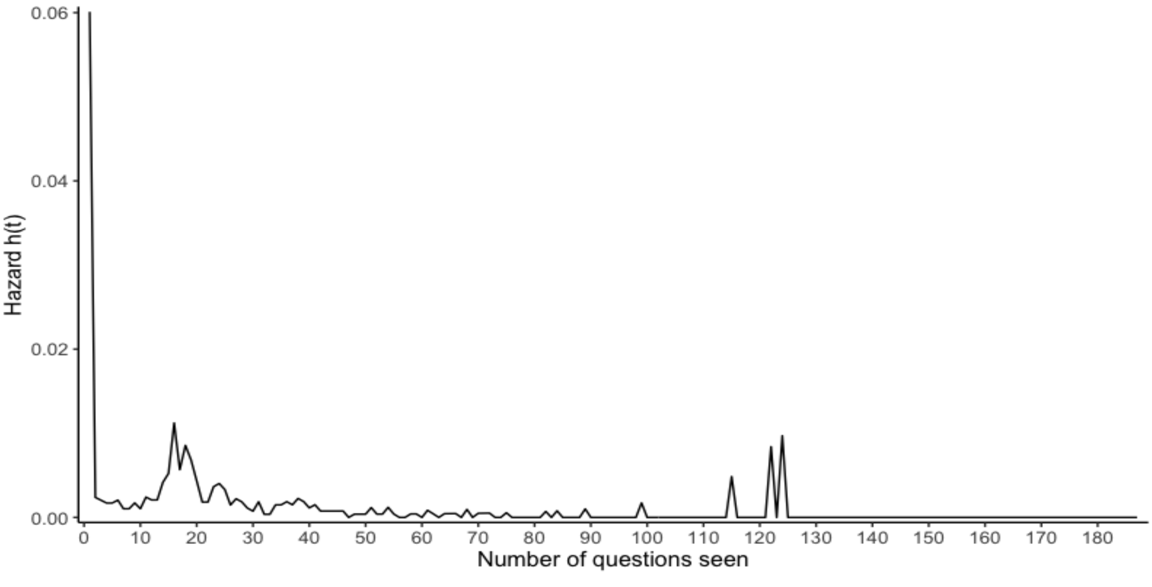


Figure 4.2. Change in the hazard of breakoffs by the number of questions seen.

As shown in Figure 4.2, the breakoff hazard is non-linearly associated with time, and there is only one change in the direction of breakoff hazard that is genuinely related to time. We therefore decided to fit all our survival models using linear and quadratic forms of time. We also conducted a sensitivity analysis by coding the time differently (See Table A.5 in Appendix A), but the quadratic time model conforms to the trend in Figure 4.2 and strikes a good balance between model interpretation, goodness-of-fit and parsimony. Thus, the quadratic time model will be reported in the following section.

4.5.2 Factors impacting breakoff and its timing

Model 1 and 2 investigate what factors impact the breakoff on the full sample. As can be seen in Table 4.2, the odds ratio of linear time (i.e. number of questions seen) of Model 1 is smaller than one, indicating that the more questions a respondent answers, the less likely she is to break off. This trend does not remain constant. The odds ratio corresponding to the quadratic time is greater than 1, so the downward breakoff likelihood flattens out to some extent as time passes by.

Model 1 also estimates the impact of different respondent demographics on survey breakoffs. Non-white respondents are 19% less likely to break off than the white peers. Students have a five-fold increase in the breakoff risk. Meanwhile, compared to respondents with a degree at the high school level or below, holders of a degree at the college level or above are about 80% less likely to break off.

Adding question characteristics and paradata to the model (i.e. Model 2) improves the overall model fit given the large AIC decrease. Compared to Model 1, the impact of student status and education are attenuated but still significant. While the odds associated with ethnicity become insignificant, age and household income become positively related to breakoffs. The odds of breakoffs for an individual who is ten years older are 10% higher. The odds of breakoff for respondents from the high household income group are 64% higher than that of peers from the low-income household.

Table 4.2. Odds ratio of logistic regression predicting breakoff (based on the full sample).

Variable	Model 1	Model 2
Intercept	0.01***	0.28***
Number of questions seen (linear)	0.92***	0.91***
Number of questions seen (quadratic)	1.0004***	1.0005***
Married (ref: no)	0.72	0.92
Male (ref: female)	1.10	1.14
Age	1.01	1.01***
Non-white (ref: white)	0.81*	0.93
Current Student (ref: no)	5.41***	2.27***
Education (ref: high school or below)		
College	0.24***	0.46***
Bachelor or above	0.20***	0.44***
Household income (ref: low)		
Middle	0.93	1.12
High	1.56	1.64**
Topic (ref: Introduction Statement)		
Demographics		0.05***
Housing		0.27***
Clothing		0.35***
Utilities		0.34***
Insurance		0.54***
Income		0.07***
Matrix question (ref: no)		1.33
Open-ended question (ref: no)		0.87
Question stem word count		0.98***
Item nonresponse rate		1.03***
Grouped (ref: Interleafed)		1.15
Mobile device (ref: non-mobile)		1.26**
Multiple sessions (ref: one session)		1.08
Survey duration (min)		0.76***
N of Respondents	3,125	3,125
N of Observations	229,816	229,816
Log Likelihood	-3,042.74	-2,380.26
AIC	6,109.48	4,812.53

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Some of the estimates in Model 2 conform to expectations. The introduction statement gives the respondent a chance to re-evaluate whether they want to continue the survey and thus is expected to associate with more breakoffs. Indeed, compared to the introduction statement, the odds of breakoffs in other topics are lower. More interestingly, when facing sensitive topics about demographics and income, respondents are approximately 95% less likely to break off compared to the introduction statement. Additionally, item nonresponse rate and mobile device are positively associated with breakoffs as expected. For every unit increase in the item nonresponse rate, the breakoff odds increase by 3%, and mobile device users would have 26% higher odds of breakoffs. Also, the more time respondents spend in the questionnaire, the less likely they will break off.

In contrast to prior studies, questions with more words are found to be associated with fewer breakoffs. More specifically, each additional word in the question stem leads to a decrease of 2% in the breakoff risk.

4.5.3 Impact of question topic and filter question format on breakoff timing

Model 3 in Table 4.3 is the result of fitting Model 2 to the restricted sample. After the restriction, the number of remaining respondents reduces from 3,128 to 2,797, of whom 188 break off. As a result, the breakoff rate declines to 6.7%.

As before, the Introduction Statement is still associated with higher breakoff odds. However, when comparing to Clothing, Insurance has a higher breakoff risk.[2] In total, there are 73 questions in the clothing block and 55 in the insurance block. The fewer questions in the insurance block and randomisation of question blocks together demonstrate that the insurance topic is genuinely associated with more breakoffs. The utilities block does not differ from the clothing block in terms of breakoff. The less-than-one odds ratio of open-ended questions in Model 4 is a surprising finding because nearly all prior studies documented that open-ended question is positively linked with survey breakoffs.

[2] We also ran a model using Insurance as the reference level for the question topic and found that both Clothing and Utilities have a lower breakoff risk compared to Insurance (See Table A.7 in Appendix A).

Table 4.3. Odds ratio of logistic regression predicting breakoff (based on the restricted sample).

Variable	**Model 3**	**Model 4**
Intercept	0.12***	0.19*
Number of questions seen (linear)	0.91***	0.89***
Number of questions seen (quadratic)	1.0006***	1.0007**
Married (ref: no)	1.10	1.10
Male (ref: female)	0.97	0.97
Age	0.998	0.998
Non-white (ref: white)	0.96	0.97
Current Student (ref: no)	0.94	0.94
Education (ref: high school or below)		
College	0.96	0.96
Bachelor or above	0.76	0.77
Household income (ref: low)		
Middle	0.96	0.96
High	1.63**	1.63**
Topic (ref: Clothing)		
Utilities	1.07	0.88
Insurance	1.74***	4.39*
Introduction Statement	2.99***	4.34
Open-ended question (ref: no)	0.42***	0.42***
Question stem word count	0.98**	0.98**
Item nonresponse rate	1.0008	1.0007
Grouped (ref: Interleafed)	1.19	0.19***
Mobile device (ref: non-mobile)	1.38**	1.39**
Multiple sessions (ref: one session)	0.95	0.95
Survey duration (min)	0.84***	0.85***
Grouped x Questions seen (linear)		1.08***
Grouped x Questions seen (quadratic)		0.999**
Utilities x Questions seen (linear)		1.01
Utilities x Questions seen (quadratic)		0.99994
Insurance x Questions seen (linear)		0.96
Insurance x Questions seen (quadratic)		1.0003
Introduction Statement x Questions seen (linear)		0.98
Introduction Statement x Questions seen (quadratic)		1.0002
N of Respondents	2,797	2,797
N of Observations	149,154	149,154
Log Likelihood	-1,269.67	-1,262.49
AIC	2,583.35	2,584.98

* p < 0.1, ** p < 0.05, *** p < 0.01

To investigate the change in time of the breakoff by question topics and filter formats, their interaction with time is included in Model 4. None of the interaction terms involving topics is significant. In contrast to Model 3, the model estimate of the grouped format on breakoffs in Model 4 becomes significant. The odds of breakoffs for respondents seeing the grouped format are only 19% of that of those answering the interleafed version. Yet, this difference varies by the number of questions respondents see. A more intuitive interpretation of the interaction effect of the grouped format and time (number of questions seen) is presented in Figure 4.3 where the fitted hazard of breakoffs generated by Model 4 is plotted against time for both grouped and interleafed formats.

As shown in Figure 4.3, when respondents see only a few questions, those receiving the interleafed format are more likely to break off. However, after approximately the 26th question, this trend is reversed; the grouped format starts to experience a higher breakoff risk. As respondents see more questions, the breakoff hazard between the two formats eventually converges (the fluctuation in both curves after the 120th question is mainly due to the small denominator in the hazard calculation). In conclusion, the grouped format can postpone the breakoff, compared to the interleafed format. However, as respondents answer more questions, the breakoff rate in the grouped format quickly catches up with that in the interleafed format.

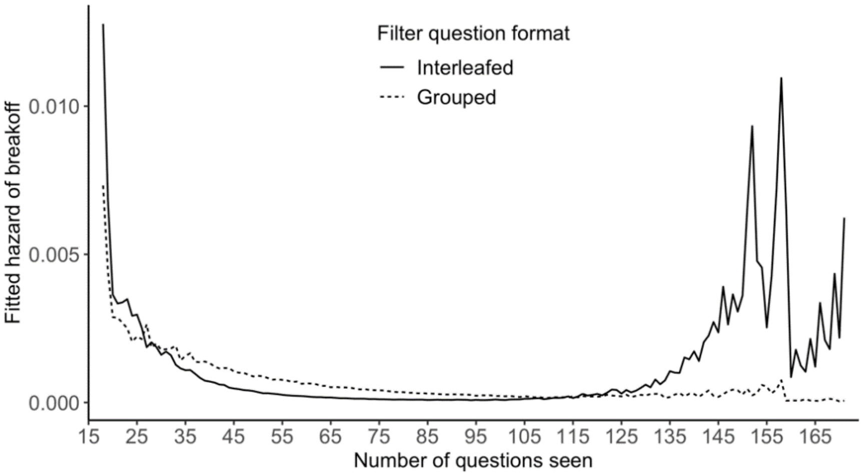


Figure 4.3. Change of the fitted breakoff hazard across time by filter question formats.

4.6 Discussion

The prevalence of the survey breakoff and the damage it can cause has led to a growing body of research into the factors causing it. This study extends this line of research by investigating two aspects of questions in particular: question topic and filter question format.

Our analysis finds two specific time points when breakoffs are more likely to happen (RQ 1). The first one is at the beginning of the survey. This finding is in accordance with previous research (Peytchev, 2009; Vehovar and Cehovin, 2014; Mittereder and West, 2021). The second timing is at the beginning of each question block where an introduction statement indicates a new set of questions. To further investigate when breakoffs are more likely to happen, two additional time-related variables are used in the analysis, namely the number of questions respondents see and survey duration. We find that the more questions respondents see the less likely they will break off. However, we remain cautious about this finding because of the possible confounding with breakoffs.

In terms of the impact of different time-varying factors on breakoffs (RQ 2), some factors are in line with prior studies. Respondents who use a mobile device to answer the survey and have a higher item nonresponse rate are more likely to break off (Wenz, 2017; Mittereder and West, 2021).

However, some predictors affect the breakoff risk in an unexpected direction. The first one is the negative relationship between word count in the question stem and breakoffs. We notice that the questions in our survey that have more words tend to be filter and follow-up questions. Most of the words in those questions are in fact repetitive. For example, every question about the price of different clothing items has the same instruction: “Round to the nearest dollar. Please include any shipping and handling charges with the cost of any item that was shipped”. When facing the repetitive content, respondents might ignore them and only read the changing parts in the question. In comparison, for questions that are shorter but not repetitive, respondents might need to read every word to understand it. This in fact makes questions with more words “shorter” while questions with fewer words “longer”. Another surprising finding is that the breakoff risk for open-ended questions is lower than that of closed ones. We suspect that this is perhaps because our survey has a large number of open-ended questions (36% of the total questions are open-ended). The extensive use of open-ended questions might make respondents accustomed to this challenging question type.

The third research question (RQ 3) investigated whether the topic of the question impacts the breakoff and its timing. Compared to the topic of clothing, the insurance-related topic suffers from a higher breakoff risk while there is no difference in the breakoff risk between clothing and utilities. Meanwhile, in comparison to the introduction statement, topics on income and demographics are relatively more sensitive but have a lower breakoff risk. Yet, the position of both blocks was not randomised, so the finding could be confounded by question order. Although some topics are associated with a higher/lower breakoff risk, we find that the relative difference between topics' effects on breakoffs remains constant regardless of how many questions respondents have seen.

The final question (RQ 4) was whether the filter question format impacts the breakoff timing. We find that the grouped format can delay the breakoff but only until respondents realise the relationship between filter question and follow-ups and gain a sense of the extra burden.

The present study also has some limitations. Firstly, the web survey analysed here has a limited amount of paradata about response behaviours (e.g., question response time). Lacking such information prohibits a more detailed analysis on the process leading to breakoffs. Also, given that some respondents broke off at Block 1 and 2, there is a possibility that these early breakoff cases might differ from those reaching Block 3 (i.e. the first of the three randomised blocks). However, our analysis includes respondents' demographic background, so we expect that this issue could be resolved to some extent. Another limitation is that the survey analysed here is a non-probability survey and skewed towards female and white individuals, making the findings less applicable to the general population. Furthermore, the topics in the survey are not fully randomly ordered, so we can only test the effect of those randomised topics on breakoffs. Moreover, respondents answering the interleafed format might learn to reduce burden by deliberately under-reporting in the filter questions. In this case, they are not shown the follow-up questions and consequently break off less often at the later stage compared to the grouped counterpart. Future research is needed to answer whether under-reporting could explain the difference in the breakoff timing between grouped and interleafed formats. Lastly, researchers can also investigate whether our finding about filter questions still hold when all filter questions are presented in a matrix format (as opposed to showing them on separate pages).

In spite of these limitations, we believe findings in this paper will be useful to survey practitioners. For example, given the fact that a large number of breakoffs happen at the introduction statement, questionnaire designers should think about ways to keep respondents engaged (e.g., placing this type of statement on the page with a substantive question or replacing this long statement with a short title about the topic). Meanwhile, findings about the insurance topic demonstrates that some question topics can impact breakoffs. Survey designers should place those topics towards the end of the questionnaire or give some motivations to the respondents in those topic blocks (e.g., emphasising the anonymity of the response). Additionally, the finding about the effect of filter question formats on breakoff timing is helpful for surveys that use filter questions extensively. For example, if the interest is in the prevalence of instances (e.g., purchase of different clothing items), the survey designer might prefer the grouped format as the postponing effect of this format would expose respondents to more filter questions. On the other hand, if the researcher cares more about the detail of the reported instance, it would be helpful to put the most important pair of filter and follow-up questions at the beginning of the interleafed format.

Chapter 5 Predicting Web Survey Breakoffs Using Machine Learning Models

Abstract

Web surveys are becoming increasingly popular but tend to have more breakoffs compared to the interviewer-administered surveys. Survey breakoffs occur when respondents quit the survey partway through. The Cox survival model is commonly used to understand patterns of breakoffs. Nevertheless, there is a trend to using more data-driven models when the purpose is prediction, such as classification machine learning models. It is unclear in the literature what are the best statistical models for predicting question-level breakoffs. Additionally, there is no consensus about the treatment of time-varying question-level predictors, such as question response time and question word count. While some researchers use the current values, others aggregate the value from the beginning of the survey. This study develops and compares both survival models and classification models along with different treatments of time-varying variables. Based on the level of agreement between the predicted and actual breakoff, we find that the Cox model and gradient boosting outperform other survival models and classification models respectively. We also find that using the values of time-varying predictors concurrent to the breakoff status is more predictive of breakoff, compared to aggregating their values from the beginning of the survey, implying that respondents' breakoff behaviour is more driven by the current response burden.

Keywords

Breakoff timing, time-varying variables, Cox model, LASSO Cox model, logistic regression, random forest, gradient boosting, support vector machine

5.1 Introduction

Web surveys have become one of the most important tools for social scientists, a trend that has been accelerated by the Covid-19 pandemic. However, running surveys on the web has some limitations, one of which is the high survey breakoff (Tourangeau, Conrad and Couper, 2013). Survey breakoff happens when the respondent starts the survey but does not complete it (Lavrakas, 2008). Consequently, the sample size available is reduced, and survey estimates can be biased when those that break off differ from those that complete the survey.

There are two main approaches to mitigating the damage of breakoffs. The first one is reactive. The differential breakoff propensity can be corrected via weighting *after* the data collection (Steinbrecher, Roßmann and Blumenstiel, 2015). The other is minimising the breakoff *during* data collection. For example, a model can continuously monitor the breakoff risk during the response process and triggers some interventions (e.g., displaying motivation messages) when the respondent is predicted to break off soon (Mittereder, 2019).

For both post-hoc correction and real-time intervention, a good prediction model of the breakoff is essential. Such a model would identify the factors strongly associated with the breakoff propensity and make weighting more effective. Also, a good prediction of the breakoff risk would help activate the intervention at the most relevant timing and potentially increase its efficiency.

The Cox survival model is widely used when studying survey breakoffs (Peytchev, 2009; Hochheimer *et al.*, 2016; Mittereder, 2019) as not breaking off implies "surviving" the response process. Previous research has shown that this traditional model can achieve a relatively satisfactory prediction accuracy (e.g., 78% for Mittereder, 2019).

However, there is a growing interest to go beyond traditional statistical models and use machine learning to improve the prediction performance even further (e.g., Lee and Lim, 2019; Spooner *et al.*, 2020). Currently, there is scant application of survival machine learning in predicting survey breakoffs. Against this backdrop, the present study will first compare the survival machine learning models with the traditional Cox model to investigate whether the former improves the performance of breakoff prediction.

Meanwhile, classification models, another class of machine learning, are widely used in survey nonresponse prediction (e.g., Kern, Klausch and Kreuter, 2019; Liu, 2020) but never applied to breakoff prediction. Models of this class usually treat each row in the data as independent. We will compare five classification models to see how they perform with regards to predicting breakoffs in the data where the rows are not independent from each other (i.e. questions in the row are clustered by respondents). These five models are: traditional and LASSO logistic regression, Random forest, Gradient boosting, and Support Vector Machine. Finally, we will compare the best performing survival model with the best performing classification model to investigate whether considering the clustered data structure by the survival model improves the breakoff prediction performance.

In addition to the statistical model used, what predictors are included and the way they are coded also play a crucial role in the model prediction performance (Kuhn and Johnson, 2013). Unlike time-constant predictors (e.g., socio-demographic variables), there is no consensus in the existing literature regarding how to treat the time-varying variables (e.g., question response time and question word count). In fact, three different treatments were used in prior studies of survey breakoffs: using the variable as it was originally coded (i.e., concurrent with the outcome) (Vehovar and Cehovin, 2014), lagged (Galesic, 2006) or accumulated (Peytchev, 2009). All these treatments are based on different assumptions regarding how the time-varying variable affects breakoffs. For example, a cumulative view of the predictors would emphasise the importance of accumulated burden in a survey while a concurrent coding would emphasise the importance of the variable where breakoff happens. While the treatment of the time-varying variables is essential for the understanding of the causal mechanisms leading to breakoffs as well as the quality of the prediction models, few researchers have explicitly tested which treatment (and the associated assumption) is better, particularly in terms of the model prediction performance. The present study will investigate this by fitting all above models with different coding of time-varying predictors.

The present study will contribute to the existing literature in two ways. First, we will help identify the model that is most predictive of breakoffs and can therefore be used both for real-time interventions and to generate weights for breakoff adjustments. Second, by investigating different ways of using time-varying predictors, this study will contribute to the theoretical debate regarding the process leading to breakoffs.

5.2 Literature review

5.2.1 Real-time intervention in web surveys

Breakoffs are common in web surveys. For example, Revilla (2017) reviewed 185 opt-in web surveys distributed through a Spanish survey company and documented a mean breakoff rate of 11.8%. In another meta-review, Liu and Wronski (2018) documented an average breakoff rate of 13% across the 25,000 non-probability web surveys implemented in SurveyMonkey.

Given the prevalence of breakoffs, survey researchers have devoted attention to studying the breakoff as a specific outcome of interest and found a number of important predictors. Some examples are using the small-screen device to answer the survey (Lugtig and Luiten, 2021) or implementing technically complicated features in the survey (Funke, Reips and Thomas, 2011). Based on the identified factors, researchers usually propose changes in designs that are prone to breakoffs, such as optimising the survey for mobile devices (Mavletova and Couper, 2015). However, most proposed solutions are reactive, meaning that survey practitioners can only make changes after data collection. In this case, proactive solutions may offer a better alternative to mitigate the damage of breakoffs. One example of proactive solutions is the real-time intervention in web surveys (Kreuter, 2017). Such systems have been implemented in surveys and proved to be useful, for example by discouraging respondents from speeding through the questionnaire (Conrad *et al.*, 2017).

The study of Mittereder (2019) is one of the few that used a proactive approach in the context of discouraging survey breakoffs. She implemented a model in a survey that continuously calculated the breakoff risk at the page level for each respondent. When the model predicted a breakoff risk that was higher than a pre-set threshold, a message appeared and highlighted the importance of completing the entire questionnaire. Prior to the survey, the sampled members were randomly assigned to three groups where the timing of displaying the message varied. The control group never had a pop-up message (i.e. no intervention) while the generic group saw the intervention message immediately at the first question (i.e. intervening irrespective of the breakoff risk). Respondents of the tailored group only received the pop-up message when the model predicted a high likelihood of quitting in the next question (i.e. intervening at the highest breakoff risk). When comparing the control group with the other two, the study found that the message was effective in lowering the breakoff rate among some specific respondents, such as students and females.

The review so far highlights the potential of the real-time intervention system in web surveys. Using such a system to effectively combat breakoffs has three pre-requisites: (1) the model can accurately predict the breakoff risk at the question level, (2) variables in the model are predictive of the imminent breakoff and (3) the interventions are effective in discouraging breakoff. Nevertheless, there is limited research in all three areas currently. The present study will explicitly tackle the first two aspects.

5.2.2 Models of web survey breakoffs

The Cox survival model is commonly used in the study of survey breakoffs (Peytchev, 2009; Hochheimer *et al.*, 2016; Mittereder and West, 2021). This model estimates the breakoff hazard which is defined as the probability of a person breaking off at a specific question given that this person has not experienced this event before (Mills, 2011). Previous research has shown that the Cox model can give a relatively satisfactory prediction performance. For instance, the study of Mittereder (2019) reviewed above reported that 78% of survey pages were correctly predicted to be (non-)break pages.

Despite the achievement of traditional statistical models, data-driven machine learning models are increasingly being used. One explanation for this change is that the traditional statistical models always include all input predictors even though some predictors are not associated with the breakoff hazard. Including many irrelevant predictors in the model can lead to overfitting and poor model interpretability (Tibshirani, 1997). In contrast, some machine learning models have a built-in feature of automatic predictor selection, which helps exclude predictors that contribute little or none to the outcome prediction (Wang, Li and Reddy, 2019). Another reason for the movement to the machine learning models is that those models are usually non-parametric, meaning that users do not need to make prior specification about the relationship between the outcome and the input predictors as the traditional statistical models require (Buskirk *et al.*, 2018). Such flexibility in the model form can reduce the chance of the model misspecification and thus has potential for improving the model prediction performance.

However, there is currently little research regarding the potential of survival machine learning models in survey research (including breakoff prediction). Against this backdrop, the first research question of this study is:

RQ1. Do survival machine learning models predict web survey breakoffs more accurately than the traditional Cox survival model?

In contrast to the scant application of survival machine learning models in survey research, classification models, another class of machine learning, are increasingly used in modelling survey nonresponse. Indeed, many survey researchers have documented that classification machine learning models could predict survey nonresponse more accurately, compared to the traditional logistic regression. Some examples of the classification models fitted in those studies are LASSO logistic regression (Signorino and Kirchner, 2018; Liu, 2020), support vector machine (Kirchner and Signorino, 2018), random forest (Buskirk, 2018) and gradient boosting (Kern, Weiß and Kolb, 2021).

However, unlike survival models, most classification models cannot handle the clustered structure in the survival data. To be more specific, many predictors of survey breakoffs (especially those question-level predictors) are time-varying, such as the question word count and question response time. To accommodate such predictors in the breakoff study, the long data format (where a row is a combination of respondents and questions) has to be used. This clustered structure (questions are clustered within respondents) means that some rows are dependent. This is different from the wide format (a row represents a unique respondent) where the classification models can easily treat each row as independent observations. Studies mentioned earlier already demonstrate that classification models can produce good prediction in the wide data, but it is unclear whether those models can still perform well when there is clustering in the data (and if so what is the best model among them). Thus, the second research question of this study is:

RQ2. What is the best classification model for predicting web breakoffs in clustered data?

If the ultimate goal of the researchers is to have a good prediction model for better survey weighting or breakoff intervention, survival models and classification models are two available options. Both classes of models approach the task of prediction differently and have its own merits, but one question remains unanswered is which of them can predict breakoff more accurately. On the one hand, accounting for the clustered structure in the data makes survival models more in line with the data generating process (Singer and Willett, 2003), which is expected to produce more accurate breakoff prediction. On the other hand, classification models focus primarily on finding the pattern between predictor values and the breakoff event and then use such a pattern to make predictions on new data. Currently, no researcher has compared survival models and classification models in the context of survival data to investigate whether respecting the clustered data structure is useful for breakoff prediction. The third research question will bridge this gap:

RQ3. Does the best performing survival model predict web survey breakoffs more accurately than the best performing classification model?

5.2.3 Predictors of web survey breakoffs

In addition to comparing different models for breakoff prediction, this study will also investigate how different types of variables affect breakoff, with a specific focus on the time-varying predictors. In the literature about breakoffs, a wide range of predictors are usually explored, but all of them can be grouped into time-constant and time-varying predictors. The former includes predictors whose values do not change throughout the questionnaire. Examples are features implemented at the survey level (e.g., the survey displays the progress bar or not) (Villar, Callegaro and Yang, 2013) and respondent characteristics, such as gender (Peytchev, 2011) and education (Blumenberg *et al.*, 2018). On the other hand, values of predictors in the time-varying group change from question to question, such as the number of characters in the question (Tijdens, 2014).

As the value of the time-constant variables always remains the same, researchers can directly include this type of variable in the model. However, for the time-varying variables, there are different approaches to include them in the modelling. Each approach is based on different assumptions about how the time-varying variable affects breakoffs.

At one extreme, some researchers accumulate the values of the time-varying variables from the start of the survey until the respondent breaks off. Supporters of this approach assume that breakoffs are caused by the gradual accumulation of burden from the beginning of the survey. An example of the variable coded in this approach is the cumulative number of questions the respondent has seen since the start of the survey (Peytchev, 2009).

Another approach of treating the time-varying variable involves no processing at all but using the original coding of the variable. In this case, researchers implicitly assume a concurrent effect of the variable: some factors are so burdensome that their presence is likely to result in a breakoff event. When including time-varying variables in the model of breakoffs, researchers use the predictor value that is concurrent with the outcome. An example is the use of a binary variable which indicates whether or not the survey page signifies more incoming questions (Vehovar and Cehovin, 2014).

Other strategies that have also been used in the past are in between the two treatments we discussed so far. One example is using the lags of the time-varying predictors. Like the accumulation approach, lagging assumes that the past event influences the breakoff likelihood. However, this approach further assumes that the recent event (as opposed to all events in the past) matters the most in breakoff prediction. For instance, Galesic (2006) used respondent's self-reported interest and burden of the previous question block (a block consists of multiple questions of the same topic) to explain their breakoffs at the next block. In this case, the lagged-one value of the time-varying variable was used.

In summary, researchers have different assumptions regarding how the time-varying variables impact breakoffs, which in turn determines the way they treat the time-varying variable in statistical models. However, little research has been conducted to explicitly compare these different treatments of the time-varying variables (and the associated underlying assumptions), especially in the context of breakoff prediction. As a result, this study will compare three different treatments of the time-varying predictors: using either only cumulative or concurrent coding and using both simultaneously. By comparing the prediction performance resulted from these different treatments, the study will answer the fourth research question:

RQ4. What is the best way to treat time-varying predictors of breakoffs in order to maximise the prediction performance?

5.3 Data

The data used in this study comes from a repeated, cross-sectional and non-probability web survey about respondents' spending on clothing, utilities bills and non-health insurance (e.g., vehicle and home insurance) (Eckman, 2021). The survey was administered to the members of the Lightspeed Panel, an opt-in web panel in the United States. Upon completing the survey, the respondents received points which can be accrued and redeemed later. Two waves were collected. The first wave was conducted between September and October 2019 while the second was collected in October 2020. The majority of the respondents in both waves are not students (around 70%), have a degree at the college level or above (70 %) and belong to the white ethnicity group (74%). The proportion of married respondents is roughly the same as the unmarried ones. Respondents in the first wave tend to be younger than those from the second wave (an average of 43 vs. 48). Although the first wave survey has more questions than the second one (196 vs. 126), the respondents of both waves spent, on average, approximately 11 minutes in the questionnaire before survey breakoff or completion.

The survey is considered appropriate to analyse for three reasons. First, it recorded the outcome of interest, namely breakoffs. Out of the 3,128 and 2,370 respondents in the first and the second waves, 520 and 403 quit the survey without completing it, resulting in a breakoff rate of around 17% for both waves. Furthermore, the survey recorded the last question respondents completed, meaning that the breakoff position is known. More importantly, the breakoff pattern of both waves is very similar. As can be seen in Figure 5.1, the highest breakoff hazard happened at the beginning of the survey for both waves. The second peak occurred after 10 to 15 questions, and questions within this range either involve sensitive topics (i.e. rent/mortgage for the dwelling), belong to matrix questions or introduce a new series of topics. The second peak occurred few questions earlier in Wave 2 simply because some questions were not asked in this wave, bringing forward those sensitive and matrix questions and the associated second peak. After the second peak, the breakoff hazard in both waves tapered off. All peaks after the 100th questions in Wave 1 were mainly due to the continually decreasing sample size involved in the breakoff hazard calculation (206 respondents remained in the survey at the 115th question compared to over 3,000 respondents when the survey started). The similar breakoff pattern across waves means that we can mix data of both waves when training and testing models.

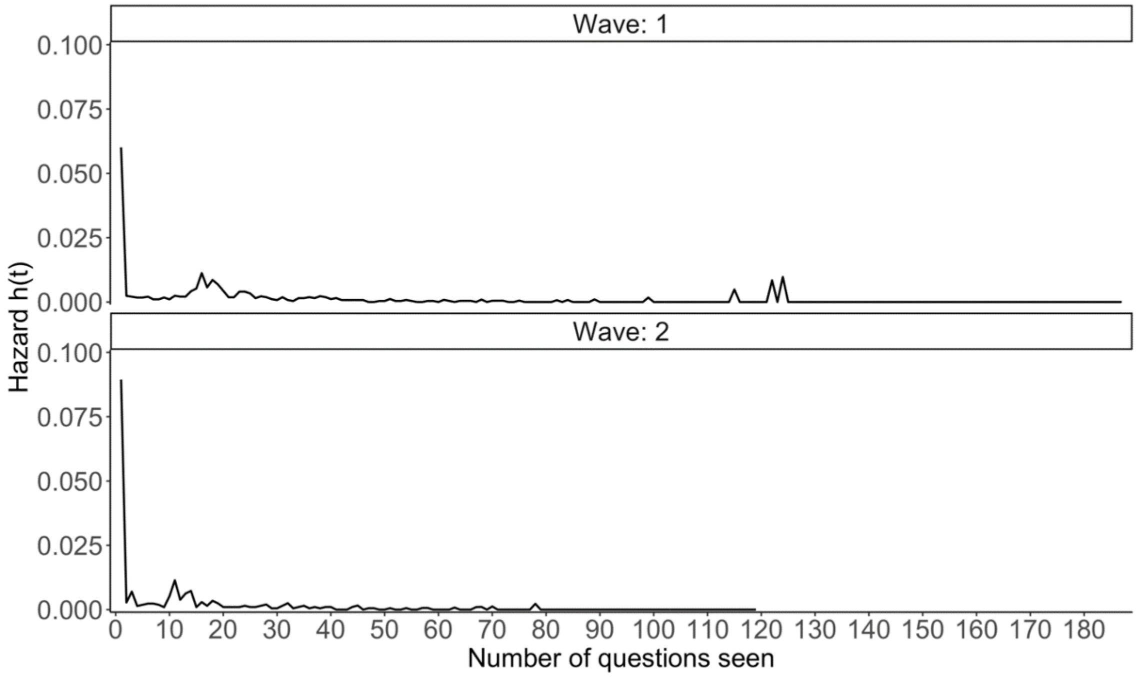


Figure 5.1. Change in the breakoff hazard by the number of questions seen for each wave.

As the survey was initially conducted to answer questions about specific survey designs, three experiments were embedded in it. The first two experiments described below were implemented in both waves, but the third experiment was only carried out in Wave 2.

The first experiment is about the filter question format. The filter question is a type of question that can trigger some follow-ups when answered positively. For example, answering "yes" to the filter question "Have you bought any jacket in the past 12 months" will activate a set of questions such as "Where did you buy it" and "How much did it cost". There are two formats for asking filter and follow-up questions, namely the grouped and interleafed format. In the grouped format, respondents see all filter questions of one particular topic (e.g., cloth items) together before moving to the follow-ups. On the other hand, individuals responding to the interleafed format are immediately exposed to the follow-ups if they answer "yes" to the filter questions. Depending on the question block, there were five to six filter questions, each of which could trigger five follow-up questions. Respondents were randomly assigned to one of the two formats.

The second experiment is related to the order of the question topic. In each wave, there were six question blocks: demographics, housing, clothing, utilities, non-health insurance and income. The first question block (Block 1) always asked respondent's demographics followed by the characteristics of their housing unit (Block 2). Questions about respondent's household income were always shown in the last question block (Block 6). Respondents' clothing purchase, utility payment and non-health insurance were randomly assigned to one of the remaining blocks (Block 3, 4 and 5). This randomisation created six possible block orderings, and respondents were randomly allocated to one of the orderings.

The third experiment (present only in Wave 2) is concerned with the order of the questions within the randomised blocks (Block 3, 4 and 5). In the first wave, the position of questions in all six blocks was fixed. In the second wave, while the order of the questions within Block 1, 2 and 6 still remained the same, the questions within Block 3, 4 and 5 were ordered in one of the two ways: (1) high-frequency to low-frequency and (2) low-frequency to high-frequency. The frequency is determined by how often the respondents selected a "yes" for the filter questions in Block 3, 4 and 6 in the first survey wave. Again, respondents were randomly assigned to one of the two groups.

All three experiment designs were crossed, so the respondents could only be in one of the 12 experimental groups in the first wave (2 filter question formats × 6 block orders) and one of the 24 experimental groups in the second wave (2 filter question formats × 6 block orders × 2 question orders).

5.4 Method

Two classes of models are fitted in this study (See Table 5.1), and both are applied to the long data (where questions are clustered by respondents) in this study to predict breakoff at the question level. The first class is the survival model, which can predict the probability of a respondent breaking off conditional upon no prior breakoff event for this respondent. The traditional Cox model and LASSO Cox belong to this class of model and are fitted in this study. Differently, models of the second class ignore the clustering of questions per respondent in the data, so they will consider each record in the data as a separate respondent and predict the probability of that record in the data having a breakoff event independently from all other records. We denote such a model class as the classification model, and five of them are fitted in this study: traditional and LASSO logistic regression, support vector machine, random forest and gradient boosting.

Table 5.1. Models fitted in this study and their tuning hyperparameters.

Model type	Hyperparameter	Description
Survival model		
Cox	- Not applicable	
LASSO Cox	- λ	- Penalty term
Classification model		
Logistic	- Not applicable	
LASSO logistic	- λ	- Penalty term
SVM	- $\varphi(\cdot)$	- Kernel function
	- C	- Penalty term
	- Sigma	- Influence of each observation has on the decision boundary (only used in radial kernel)
	- d	- Polynomial degree (only used in polynomial kernel)
Random forest	- trees	- Number of trees
	- mtry	- Number of predictors to consider for a split
	- min_n	- Minimum number of cases in a node for a split to continue
Gradient boosting	- trees	- Number of trees
	- mtry	- Number of predictors to consider for a split
	- tree_depth	- How many splits a tree can have
	- learn_rate	- Learning rate
	- min_n	- Minimum number of cases in a node for a split to continue

5.4.1 Traditional Cox model

As mentioned earlier, predicting survey breakoffs is essentially a survival problem, and the traditional Cox model is designed to handle the survival data. It can explain whether breakoff will happen and if so its timing (Singer and Willett, 2003). We code the timing as the number of questions seen by respondents up to the point where they broke off or completed the survey, which is an approach commonly used in the prior literature (Peytchev, 2009; Mittereder and West, 2021).

As shown in its equation below, the breakoff hazard at a specific time t is modelled as the product of the baseline hazard $h_0(t)$ (i.e. the breakoff risk when all predictor values are zero or at the reference level) and the predictor's multiplicative effect β on the breakoff hazard.

$$h(t) = h_0(t) \cdot e^{\beta x}$$

To estimate the coefficients β, the partial likelihood is used. For computational efficiency, the negative log transformation of the partial likelihood is actually estimated, and the algorithm will identify the combination of parameters that minimises the negative log likelihood. Given the popularity of the Cox survival model in estimating breakoff risks, it is used as a benchmark for other survival models fitted in this study.

5.4.2 LASSO Cox

LASSO Cox builds on the traditional Cox survival model by adding a penalty term to the negative log likelihood estimate to penalise those models that have many parameters (Tibshirani, 1997). The penalty term λ is a non-negative hyperparameter and controls the degree of penalisation. When λ is zero, no penalisation is imposed, and the fitted model is the same as the traditional Cox model. The larger the λ the more estimated model coefficients are shrunk towards zero. Some of the coefficients will be equal to zero when λ is large enough and they will be excluded from the final model. This automatic predictor selection is especially useful for complex models where there are a large number of predictors. The LASSO Cox model is used in this study to see whether a simpler model would outperform the full-size Cox survival model.

5.4.3 Support vector machine

Like LASSO Cox model, the support vector machine (SVM) also uses a penalty term and aims to minimise a defined function. Despite this similarity, SVM operates very differently. It tries to find a hyperplane in the high-dimensional space defined by the number of predictors such that the breakoff and non-breakoff respondents, represented as points in the space, can be linearly separated (Rhys, 2020). Given that there might be many hyperplanes that could perfectly separate the two classes, the best one must satisfy two criteria. It should be farthest from the points of both classes, *and* points of one class (y = 1) lie above this hyperplane while points of the other class (y = -1) fall below it.

However, two classes (in this case respondents who complete and break off the survey) are rarely perfectly separable by a linear plane in practice. SVM addresses this issue in two ways. Firstly, users can extend the space by transforming the predictors. The transformation is equivalent to adding more dimensions to the data, which makes the data linearly separable in the enlarged dimensions. The function for the transformation is called the kernel $\varphi(\cdot)$. Some commonly used kernels are linear kernel (i.e. no transformation), polynomial kernel (e.g., quadratic) and radial kernel. The second approach for finding a linear separating hyperplane is allowing some misclassification (i.e. the respondent is predicted by the model to be a breakoff case though the reverse is true). The extent of allowed misclassification is controlled by a non-negative penalty term C. Larger C will impose more penalty on the misclassification and push the fitting algorithm to work harder to produce more correct classification.

The kernel function $\varphi(\cdot)$ and the penalty value C are two of the hyperparameters in SVM. Using the hyperplane with the tuned hyperparameters, the respondents can be classified as breakoff or non-breakoff cases depending on which side of the hyperplane they are predicted to fall into.

5.4.4 Random forest

Both random forest and the boosting require fitting multiple decision trees. Based on the breakoff status, the decision tree recursively partitions the respondents into two child nodes using one of the input predictors. As such, respondents within the same node become more homogeneous in terms of their breakoff status while respondents between nodes are more dissimilar (Buskirk *et al.*, 2018). Gini index quantifies the heterogeneity of a node based on the proportion of breakoff and non-breakoff respondents in the node. The higher the Gini index the less homogeneous (or more heterogeneous) the cases in the node are. By comparing the Gini index of a parent node with that of the child nodes resulted from using different predictors to split the tree, the decision tree will choose the predictor which can make the largest reduction in the Gini index compared to the parent node. The splitting process will stop according to some user-defined criteria such as the minimum number of respondents required in a node for making a further split (Rhys, 2020).

Although the random forest and boosting are based on the decision tree, the way they develop a single tree slightly deviates from the fitting process described above. The random forest incorporates two randomised elements into the tree growth. Both randomisations reduce the chance of overfitting and are the two main tuning hyperparameters in random forest. Firstly, rather than using the same data to train the tree model, the random forest will randomly draw *B* bootstrapped samples from the original data where each of the bootstrapped sample is as large as the original dataset (James *et al.*, 2013). Then, the decision tree is independently developed on each of the *B* bootstrapped samples. The second randomisation happens when selecting predictors to make a split. Instead of considering all input predictors as the candidate for making a split, only a random subset of the predictors is considered in the random forest (James *et al.*, 2013).

When using random forest to predict the breakoff status, each tree in the forest produces its own prediction (i.e. breakoff or non-breakoff), then the model makes the final prediction by choosing the most frequent predicted breakoff status across all the trees.

Random forest is used in this study for two reasons. First, this model requires little data pre-processing in contrast to others machine learning models, and it can automatically handle some complex model structures (e.g., interactions). Also, most users are familiar with the tree structure, thereby facilitating the interpretability to a degree. All in all, we expect the random forest to strike a good balance between model prediction performance and interpretability.

5.4.5 Gradient boosting

Like the random forest, gradient boosting also involves fitting multiple trees. However, the two models differ in three main aspects, namely (1) how the fitting algorithm starts, (2) whether the growth of the subsequent tree depends on the preceding trees, and (3) what is the dependent variable.

Specifically, gradient boosting begins by assigning every respondent the same constant (e.g., the average breakoff hazard calculated from the data). The prediction error for each respondent is then simply calculated as the difference between this assigned breakoff hazard and the observed breakoff status (i.e. 0/1). A decision tree is then fitted using the prediction error as the dependent variable. Once this tree is developed, it is combined with the initial average breakoff hazard to make new predictions about the breakoff hazard for all respondents. Taking the difference between the new prediction and the observed breakoff status will lead to a new set of prediction errors, which another tree will proceed to model. This three-step cycle (i.e. fitting trees on the prediction errors, making new predictions by combining the initial average value and all trees fitted so far, and calculating prediction errors) will continue until some user-defined conditions are reached, such as the maximum number of tress allowed.

In gradient boosting, every subsequent tree gradually learns the prediction mistake made by previous ones and improves upon it. Existing literature has shown that the final model from this gradual model development tends to perform better (James *et al.*, 2013). There are many hyperparameters to be tuned in gradient boosting, such as the tree depth (how many splits are performed in a tree) and the learning rate (how quickly the tree learns from the previous mistake). Gradient boosting, SVM and random forest ignore the clustering of the data and it is unclear whether they can improve the breakoff prediction than the two survival models described earlier.

5.5 Analysis plan

To develop and evaluate all the models, we combine both waves of the Lightspeed Panel data and draw a stratified random sample (stratified by wave and breakoff status) with a proportion of 75% of the total rows in the data. These 75% selected rows are used as the training data, and the remaining data are the testing data.[3] All models are built and tuned using the training data, after which the models are applied to the unseen testing data to predict breakoffs at the question level. The level of agreement between the predicted breakoff from the model and the true breakoff status from the testing data forms the basis for evaluating the model performance.

Because the training data are in the long format and many questions do not have a breakoff event (99.75% vs. 0.25%), a class imbalance problem exists. To solve its negative impact on the utility of classification models, we down sample the training data when building the classification models so that the ratio between breakoff and non-breakoff questions is 1:1 (Kuhn and Johnson, 2013). Given that the survival models can handle the clustered data structure, no class balancing is applied to the training data when fitting such models. For

[3] The sampling was conducted at the respondent level, meaning that once a respondent was selected to be in the training data all question-level data of the selected person were included in the training data.

survival and classification models, they are evaluated on the same testing data (where no class balancing is carried out). In total, the original training data for developing the survival models have 274,658 rows compared to 1,390 rows in the down-sampled training data for developing the classification models. The testing data have 92,952 rows.

The data suffer from some levels of missing demographics because some respondents quit the survey before answering those questions. The level of missing demographics varies by the variables but ranges from 1% to 12% (See Table B.1 and Table B.2 in Appendix B). To include those respondents in the model development, we decided to code the missingness as an explicit category in the demographic variables. This decision is likely to artificially increase the prediction performance of demographic predictors. This is because respondents who broke off prior to the demographic questions will always have the missingness category in those variables. However, it is still necessary to code the missingness explicitly because it minimises the sample loss during the model development. The sample size plays an important role in the development and tuning of machine learning models, which partitions the data into multiple small subsets. An insufficient number of breakoff cases in the partitioned data might prevent the data-driven machine learning algorithm from reaching its optimum performance. Additionally, in most surveys, demographic information is usually unknown a priori and can be only collected from the questions in the survey. Thus, when respondents do not answer questions on their demographic background, it leads to missing demographics, and coding the missingness explicitly is one of the solutions that researchers can adopt in practice. Overall, the benefits of allowing the model fitting algorithm to function properly and developing predictive models for the real-world situation outweigh the risks associated with coding the missing demographics explicitly.

To evaluate the performance of the survival models, the concordance index (C-index) is used. This metric quantifies the proportion of respondent pairs in which the breakoff case has a higher predicted breakoff risk (Harrell, Lee and Mark, 1996). The C-index with a value of 0.5 indicates that the model is just as good as the random guess while a value of one means that the model can perfectly distinguish breakoff respondents from those who complete the survey. Because the C-index is specifically designed to work with clustered and imbalanced data (which is the case for the testing data in this study), it will be used primarily to quantify the prediction performance of survival models in this study.

In contrast, there is no guidance on how to evaluate the prediction performance of the classification model when it is applied to clustered and imbalanced data. Therefore, five metrics (Accuracy, Sensitivity, Specificity, Precision and AUC) are used to evaluate the performance of the classification models from different aspects (Kuhn and Johnson, 2013). Accuracy measures the proportion of correctly predicted breakoff and non-breakoff cases out of all available records. Sensitivity and Specificity quantify the proportion of correctly predicted breakoff/non-breakoff cases out of all actual breakoffs/non-breakoff records respectively. Precision is a metric about the proportion of actual breakoff cases out of all the predicted breakoff cases. Differently, AUC (area under the receiving operator curve) is an aggregate metric, which summarises the model performance across different combinations of the true positive rate (also called Sensitivity) and false positive rate (1 - Specificity). All five metrics range between zero and one, and the larger the value the better the model prediction performance is.

As with the development of machine learning models in other studies, we also tune some model hyperparameters. All these hyperparameters are listed in Table 5.1. We use a random grid search in the tuning where 100 combinations of the hyperparameter values were randomly tried for each model (Kuhn and Johnson, 2013). Tuning is performed using the five-fold cross-validation in the training data. The best hyperparameter value identified via the cross-validation is used when re-fitting the corresponding model on the entire training data. In the end, there are 28 models (7 model types × 4 predictor groups). These 28 models will be evaluated using the testing data.

To answer RQ 1 about whether LASSO Cox is better than the traditional Cox model, their C-index values are compared. For RQ2 (what is the best classification models for predicting web survey breakoffs), Accuracy, Sensitivity, Specificity, Precision and AUC are compared among classification models. These five metrics will be used again when answering RQ3 where the outperforming survival model from RQ1 and the outperforming classification model from RQ2 are compared. This is because both classes of the model can generate these five metrics, which ensures that the comparison is on an equal footing.

Unlike the between-model comparison in the first three research questions, answering RQ4 (how different predictor groups affect the prediction performance) will involve within-model comparison. More specifically, the comparison of evaluation metrics is performed across four predictor groups (See Table 5.2), which is repeated for each of the seven model types mentioned earlier. The prediction performance of the models that only have respondents' demographic background is treated as the baseline. Time-varying question-level predictors are coded in two ways, namely concurrent and cumulative. Predictors coded in the former way need no pre-processing and are used in the model directly. In contrast, the cumulative coding will aggregate the values of the predictors question by question. For instance, the predictor about the open-ended question will be binary (i.e. whether or not a specific question is open-ended) in the concurrent coding, but the cumulative coding will record how many open-ended questions the respondents have seen so far. Comparing the baseline prediction performance with that of the two models that respectively include the concurrent and cumulative time-varying predictors, we can investigate whether the time-varying predictor groups are more predictive of breakoff and which coding is better. Finally, respondents' demographic information and both time-varying treatments are simultaneously put into the model to see whether using all available predictors improves prediction performance.

Table 5.2. Predictors used in this study.

Predictor group	Predictors
Demographics	Age, education*, ethnicity, student status*, marital status*, filter question format, and question order*
Concurrent	Responding device, item missing*, matrix question, open-ended question, question topic, and question word count, filter question format, and question order*
Cumulative	Item missing (cumulative)*, matrix question (cumulative), open-ended question (cumulative), question topic (cumulative), and question word count (cumulative), and number of times respondents logged into the web survey (cumulative), filter question format, and question order*
All combined	All predictors above

* Its main effect and interaction with time are both included in the model.

Regardless of which predictor group is used, variables about the two experiments (i.e. filter question formats and question orders of the high-low frequency) are always included in the models as control variables. The variable about the block orders is not included because its information is already represented by the time-varying variable about the question topic. Given that some predictors violate the proportionate hazard assumption of the traditional Cox model, we interact those violating variables with time (i.e. number of questions seen) in all models fitted in this study. Those variables are marked with an asterisk in Table 5.2. The descriptive summary for all predictors in this study and their coding are provided in Table B.1 and Table B.2 of Appendix B.

5.6 Results

5.6.1 Comparing survival models

The C-index in Table 5.3 shows that the traditional Cox model tends to perform better than the LASSO Cox in predicting breakoffs. Indeed, the traditional Cox survival model can achieve a C-index between 0.68 and 0.85, compared to 0.5 to 0.78 in LASSO Cox.

Table 5.3. Prediction performance of survival models applied to testing data.

Model type	**Predictor group**	**Hyperparameter**	**C-index**
Cox	Demographics	-	0.78
	Concurrent	-	0.74
	Cumulative	-	0.68
	All combined	-	0.85
LASSO Cox	Demographics	$\lambda = 0.00988$	0.78
	Concurrent	$\lambda = 0.00040$	0.69
	Cumulative	$\lambda = 2.72$	0.50
	All combined	$\lambda = 0.00988$	0.78

Looking at the predictor groups, the traditional Cox survival model also generates a higher C-index than the LASSO Cox in three predictor groups. The only exception is when the respondents' demographic information is used as the predictors alone in the model where both traditional and LASSO Cox models produce the same C-index. This finding can be explained by the way demographics are coded in this study. As mentioned earlier, explicitly coded missingness in the demographic variables can artificially increase the correlation between this particular category and the breakoff outcome. As a result, even though both models include different numbers of demographic predictors (the traditional Cox survival model uses all input demographic predictors while the LASSO Cox has only a subset of them due to the penalisation), they can simply predict breakoff using the missingness (rather than other categories in the demographic variables). Indeed, a close examination of the model result shows that the missingness in demographics makes a dominant influence on the prediction of breakoff in both models. Because of this, both models fitted using only demographics make the same prediction and produce the same C-index.

From the perspective of the penalty term, it can also be concluded that the traditional Cox model is preferred over the LASSO Cox for predicting breakoffs. To be more specific, the best penalty values in LASSO Cox are very close to zero (except for the LASSO Cox fitted using only cumulative time-varying predictors), which implies that that there is little need for penalisation. The large λ value in the LASSO Cox using the cumulative coding is related to the non-convergence issue during the model development. To be specific, some λ values led to non-converged models in some folds during the cross validation. The reported λ value of 2.72 was the smallest one among those that were trialled *and* led to a converged LASSO model in *all* five folds of the cross-validated data. However, using this λ value, none of the input cumulative time-varying variables is retained in final LASSO Cox. This implies that it might be ineffective to use only the cumulative time-varying variables when predicting breakoffs.

5.6.2 Comparing classification models

The next between-model comparison is conducted among classification models. Different metrics to evaluate the prediction performance of those models are presented in Table 5.4. Looking at the range of all metrics, most of them (except for Precision) are above 0.75, indicating that all the classification models achieve a good prediction performance. However, the precision is very low (0.01 for the majority of the models), meaning that very few breakoffs predicted by the models are in fact breakoffs and the models raise too many false alarms about breakoff.

Table 5.4. Prediction performance of classification models applied to testing data.

Model type	**Predictor group**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
Logistic	Demographic	0.85	0.62	0.85	0.01	0.85
	Concurrent	0.81	0.73	0.81	0.01	0.87
	Cumulative	0.80	0.70	0.80	0.01	0.84
	All combined	0.84	0.78	0.84	0.01	0.91
	Average	0.83	0.71	0.83	0.01	0.87
LASSO logistic	Demographic	0.85	0.59	0.85	0.01	0.85
	Concurrent	0.81	0.73	0.81	0.01	0.87
	Cumulative	0.77	0.73	0.77	0.01	0.85
	All combined	0.86	0.76	0.86	0.01	0.91
	Average	0.82	0.70	0.82	0.01	0.87
Random forest	Demographic	0.85	0.75	0.85	0.01	0.89
	Concurrent	0.82	0.76	0.82	0.01	0.87
	Cumulative	0.83	0.72	0.83	0.01	0.87
	All combined	0.86	0.76	0.86	0.01	0.91
	Average	0.84	0.75	0.84	0.01	**0.89**
Gradient boosting	Demographic	0.82	0.79	0.82	0.01	0.90
	Concurrent	0.82	0.77	0.82	0.01	0.88
	Cumulative	0.81	0.72	0.82	0.01	0.86
	All combined	0.85	0.77	0.85	0.01	0.91
	Average	0.83	**0.77**	0.83	0.01	**0.89**
SVM	Demographic	0.99	0.55	0.99	0.14	0.84
	Concurrent	0.83	0.69	0.83	0.01	0.86
	Cumulative	0.79	0.69	0.79	0.01	0.84
	All combined	0.83	0.80	0.83	0.01	0.91
	Average	**0.86**	0.68	**0.86**	**0.04**	0.86

To facilitate the model comparison, the average of different metrics across the four predictor groups is calculated for each model (highlighted in grey). The highest average metric value across different models is in bold. As can be seen, gradient boosting gives the best prediction performance from the perspective of Sensitivity and AUC. In terms of Accuracy, Specificity and Precision, SVM outperforms the other models.

However, SVM excels in those three metrics mainly because of its performance in the predictor group of demographics alone. More specifically, fitting SVM with only demographics leads to 0.99 in both Accuracy and Specificity, which greatly increases the average value of these two metrics for SVM. However, the Sensitivity of SVM (demographics only) is 0.55, meaning that only 55% of actual breakoffs are captured in the model prediction. Thus, it can be concluded that SVM achieves an overall good prediction performance by simply predicting cases to be non-breakoff most of the time, which comes at the expense of missing many actual breakoffs. Given that the models in this study are developed to predict breakoffs and subsequently trigger real-time interventions, it is important to capture actual breakoffs to a large extent. The fact that SVM has a very low Sensitivity among the five models means that this model might not be suitable for the prediction task.

Looking at the traditional logistic regression, its prediction performance is close to that of some machine learning models. Indeed, both the traditional and LASSO logistic regression models give a similar performance across the five metrics for model evaluation. This corroborates one of the findings when comparing traditional and LASSO Cox models (i.e. there is perhaps no need for penalisation). Nonetheless, the traditional logistic regression performs less well in Sensitivity and AUC, in comparison to the random forest and gradient boosting. This means that the flexibility of the ensemble models can improve upon the parametric logistic regression and therefore should be used as the predictive model.

The performance difference between the random forest and gradient boosting is less noticeable because both models generate the highest AUC and similar Accuracy, Specificity and Precision. However, compared to the random forest, gradient boosting produces a higher Sensitivity. Also, given that each tree in gradient boosting focuses on correcting the mistakes made by the previous trees, this model has the potential of capturing more complex patterns in the data. Overall, we conclude that gradient boosting outperforms the other classification models fitted in this study. The best hyperparameter values for all models are presented in Table B.3 in Appendix B.

5.6.3 Comparing the best performing survival and classification models

Table 5.5 shows the comparison between the traditional Cox survival model and the gradient boosting classification model (both outperformed other models among their own class). As shown in the table, both models perform equally well as they have similar values in different evaluation metrics. Looking at the AUC, a metric that takes into account both the true positive rate and false positive rate, gradient boosting is slightly better. Given that gradient boosting has fewer model assumptions (in contrast to the proportional hazard assumption in the traditional Cox survival model), we therefore conclude that extra consideration of the clustering data structure by the survival model might be unnecessary as it does not translate into a significant improvement in the performance of breakoff prediction.

Table 5.5. Prediction performance of the best survival model and the best classification model applied to testing data.

Model type	Predictor group	Accuracy	Sensitivity	Specificity	Precision	AUC
Cox	Demographic	0.82	0.79	0.82	0.01	0.88
	Concurrent	0.82	0.75	0.82	0.01	0.86
	Cumulative	0.79	0.77	0.79	0.01	0.85
	All combined	0.87	0.78	0.87	0.01	0.89
Gradient boosting	Demographic	0.82	0.79	0.82	0.01	0.90
	Concurrent	0.82	0.77	0.82	0.01	0.88
	Cumulative	0.81	0.72	0.82	0.01	0.86
	All combined	0.85	0.77	0.85	0.01	0.91

5.6.4 Comparing predictor groups

When comparing the prediction performance between different groups of predictors (i.e. within-model comparison), Table 5.3 (for survival models) and Table 5.4 (for classification models) together show that using all available predictors frequently results in the highest value in different evaluation metrics. Meanwhile, using only the cumulative predictors leads to the worst prediction performance most of the time. Therefore, the concurrent coding seems to be more predictive of breakoff than the cumulative coding. However, the performance ranking for demographic and concurrent predictors is less clear as it varies by both the models and the evaluation metrics. Overall, using demographics seems to be more predictive of breakoff than using concurrent predictors.

To understand what variables in the best performing model (i.e. gradient boosting fitted with all the variables) contribute the most to the prediction of breakoff, the variable importance plot is presented in Figure 5.2. The x-axis shows the importance score, which quantifies the extent to which the model replies on the variable when making predictions. A larger importance score indicates that the variable is more important for breakoff prediction. For each variable, its importance score is calculated by summing up the change in the Gini index across all the trees where that variable is used to make a split and then taking the average (Hastie, Tibshirani and Friedman, 2009). The 10 most important variables are presented on the y-axis.

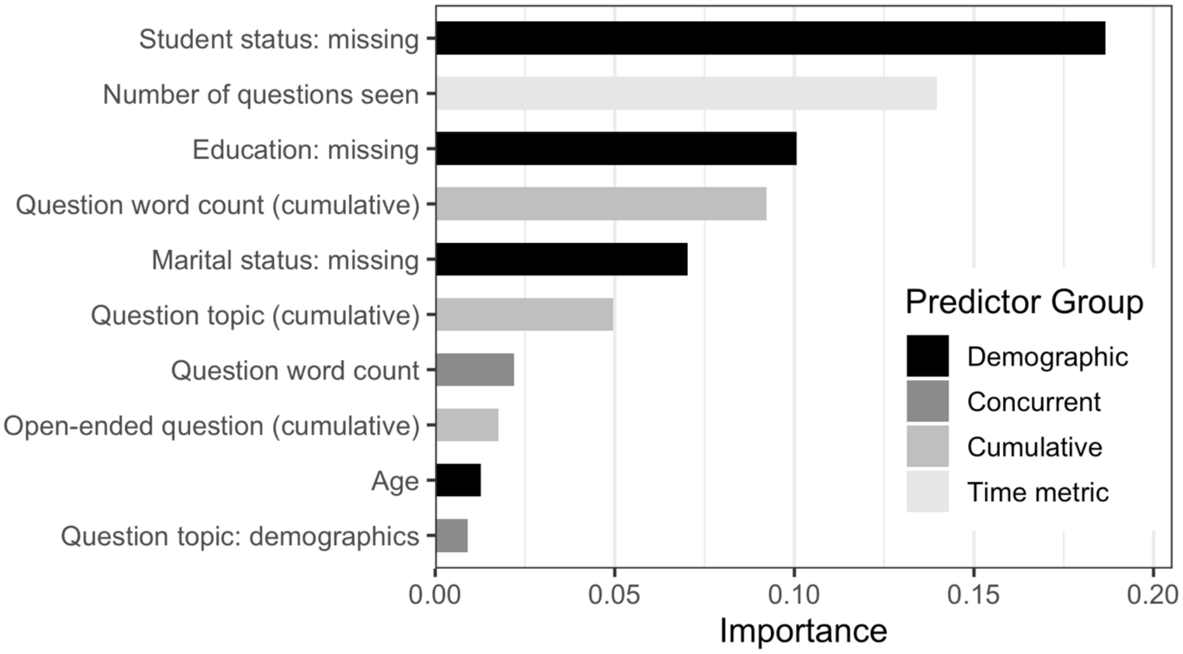


Figure 5.2. Variable importance plot for the gradient boosting (only the top 10 most important variables are shown).

As shown in Figure 5.2, variables of different predictor groups are in this top 10 list, meaning that all types of variables can contribute to the breakoff prediction. Meanwhile, three of the four demographic variables in this top 10 list are related to the category of coded missingness. This finding is not surprising because, as explained earlier, coding the missingness in the demographics as an explicit category artificially increases the association between this category and the breakoff status. Another unsurprising predictor in this top 10 list is the time (represented by the number questions seen). It has been found to be associated with breakoffs by many researchers such as Hoerger (2010).

It might be surprising to see that some cumulative time-varying variables not only exist in this top 10 list but also rank higher than the concurrent counterparts, especially considering the earlier finding that using only cumulative time-varying variables often produces the worst prediction performance. However, readers should be reminded of the concept of ecological fallacy. It happens when researchers draw wrong conclusions about individuals using findings from the groups to which the individuals belong (Brewer and Venaik, 2014). In this study, it means researchers are making conclusion about the importance of individual variables using findings from the predictor groups to which the individual variables belong. To be more specific, the earlier comparison was conducted at the level of predictor groups, and the conclusion was that *in general* the current burden is more influential on the breakoff event than the cumulative burden. This group-level finding does not conflict with the finding that *some* specific cumulative time-varying variables can be more important than the concurrent time-varying counterparts in the breakoff prediction, which is the conclusion from the present predictor-level comparison.

There are three cumulative time-varying variables that are important for the breakoff prediction, namely the cumulative number of the question word count, question topic, and open-ended question. Meanwhile, the concurrent coding of the question word count and the question topic is also found to be important for the breakoff prediction. Given that both cumulative and concurrent coding schemes of these variables are present in the top 10 list, the predictive model should include these two variables coded in both ways.

5.7 Discussion

Researchers have discovered that post-collection weighting and real-time intervention are two promising methods to mitigate the impact of breakoffs. The present study extends this line of research by comparing what models are more predictive of breakoffs and thus help derive better weighting and trigger the intervention at the most relevant timing. Also, this study bridges a previous research gap by investigating what variables and how they should be used to maximise the performance of question-level breakoff prediction.

By comparing the C-index of traditional and LASSO Cox models for the survival models, we find that the LASSO Cox does not outperform the traditional Cox (RQ1). This finding is in line with the result from the same comparison but in the medical field (e.g., Lee and Lim, 2019; Spooner *et al.*, 2020). Altogether, it implies that the Cox model is perhaps already flexible enough to create good prediction in the survival context. This can partly be explained by the semi-parametric nature of the traditional Cox model where users do not need to specify the baseline hazard, which reduces the chance of model misspecification. Another possible explanation is that there might not be a large number of predictors in our study to allow the automatic feature selection of LASSO Cox to function properly.

Among the five classification models fitted for the binary classification of breakoffs, we found that gradient boosting gives the best prediction performance overall (RQ2). This model has been found to be the 'winner' in many machine learning comparisons (Bojer and Meldgaard, 2021). The commonly cited reason is that this model focuses on correcting for the prediction errors made by the models in previous iterations (James *et al.*, 2013). Over time, the model will make fewer prediction errors and thus result in better prediction performance.

The most interesting between-model comparison is between the outperforming survival and classification models (RQ3). In our study, it is between the traditional Cox survival model and gradient boosting. We found that gradient boosting outperforms the traditional Cox model in terms of AUC. This is interesting because researchers who choose to fit the traditional Cox survival model to the survival data assume that taking account of the clustered data structure will provide more validity to the model. However, our study reveals that gradient boosting, while ignoring the clustered data structure, can still correctly predict many survey questions as (non-)breakoff questions. Equally importantly, gradient boosting can achieve such a good prediction performance but not at the expense of model interpretability. Users of the gradient boosting model can still learn what predictors are most importance for predicting breakoffs using the variable importance plot or investigate how a specific predictor impacts the breakoff risk using the partial dependence plot (Christoph, 2019). We therefore recommend practitioners to deploy the gradient boosting model if their goal is to use real-time interventions to combat web survey breakoffs.

When comparing the prediction performance between predictor types (RQ4), we found that using all available predictors always gives the best prediction performance across different metrics. Given that both concurrent and cumulative predictors contribute to the prediction, we can conclude that both current burden and the burden accumulated since the start of the survey can cause survey breakoffs. However, when only using time-varying predictors, the concurrent coding is better than the cumulative counterpart. This implies that respondents' decision to continue or quit the survey is more driven by the question they are seeing in the moment. We remain cautious about the finding that demographics alone is more predictive of breakoffs than concurrent predictors. This is because we coded the missing demographic information explicitly as a category in the predictor, so respondents who broke off before seeing the demographic questions will always have missing data in demographic-related predictors. Our coding could therefore artificially increase the predictive performance of demographic predictors. Future research can easily solve this issue by using demographics from the sampling frame.

Our study has limitations as well. To begin with, we can only fit one survival machine learning model (i.e. LASSO Cox). This is mainly because existing software packages for fitting survival machine learning models are not mature enough to handle the long data format. Even though the fitting algorithm for LASSO Cox is designed to work with the long data, estimating the LASSO Cox with a lambda value of zero (which in theory is equivalent to fitting a traditional Cox model) led to a non-convergence result while the traditional Cox model converged successfully on the same data. Secondly, we derived most time-varying predictors from the question characteristics (e.g., open-ended, number of words), and there is a limited number of time-varying predictors about respondents' behaviours (e.g., the number of survey logins). Prior research has demonstrated that response behaviours can shed more light on the process leading to breakoffs (Mittereder and West, 2021). Future research can explore how different coding of the behaviour-related predictors affects the prediction performance. Lastly, this study cannot investigate whether the lagged version of the time-varying predictors is more predictive of breakoff compared to other coding approaches. This is because creating lags will result in the first few rows of each respondent having missing data in the time-varying predictors. Because some breakoffs happened at the first question, breakoff cases with missing time-varying predictors will be removed and the sample size for model building will be noticeably reduced.

Despite the limitations, our findings can still provide some practical implications. When predicting breakoffs, gradient boosting might be the best candidate model, and concurrent and cumulative coding of the time-varying variables should be simultaneously included as predictors in the model. Future research can extend our study by looking at whether the improved prediction performance leads to better survey weighting and more efficient breakoff interventions.

Chapter 6 Comparing Different Methods of Compensating for Survey Breakoffs

Abstract

Web surveys are popular given their low cost and short turnaround time but tend to have more breakoffs compared to interviewer-administered surveys. Survey breakoffs occur when respondents quit the survey partway through. It causes missing data, which further results in reduced sample size for analysis, lower statistical power and may lead to biased survey estimates. Few prior studies investigated how to mitigate the breakoff bias in survey estimates. This study develops a simulation where the breakoff rate and the cause of breakoffs are manipulated. Four methods are then applied to compensate for the breakoff: (1) ignoring breakoff completely in the data analysis, (2) classifying breakoff as survey nonresponse and then weighting the data using nonresponse propensity, (3) treating breakoff as a special survey outcome and weighting the data by nonresponse and breakoff propensity combined, and (4) multiple imputation. We find that multiple imputation compensates for the breakoff bias slightly better than other methods and using breakoff weighting is the least preferred option. Also, the breakoff mechanism is more influential on the effectiveness of the method for breakoff compensation than the breakoff rate. Additionally, none of the methods employed here can correct for the breakoff bias when the data are Missing Not At Random. Based on the findings, we suggest that the breakoff event should be accounted for using multiple imputation when the data are Missing Completely At Random or Missing At Random.

6.1 Introduction

Online probability surveys have been used widely to support policy making, such as investigating the impact of Covid-19 on different ethnicity groups (Morales, Morales and Beltran, 2021) or understanding citizens' concerns about country's energy security (Caferra, Colasante and Morone, 2021). Having a representative survey sample is key to providing high-quality evidence for decision-making. However, many factors can affect survey representativeness. One of these factors is survey breakoff, which occurs when respondents start the survey but do not complete it (Tourangeau, Conrad and Couper, 2013).

The occurrence of breakoff leads to a smaller sample size, lower statistical power and may result in biased survey estimates (Steinbrecher, Roßmann and Blumenstiel, 2015). Even when the goal of data collection is to make causal inference among the participants in a randomised control trial (i.e. rather than generalising findings to the population), participants with certain characteristics might break off more often in the trial, which can bias the finding in the trial (Leon *et al.*, 2006).

Given the negative effects of breakoffs, many researchers have studied this survey outcome. The majority of them focused on exploring what factors impact breakoffs and proposed some solutions, such as embracing mobile-friendly designs (Mavletova and Couper, 2015). Nonetheless, no matter how successful those solutions are, breakoff is unlikely to be eliminated. It is therefore important to use post-collection methods to compensate for this type of nonresponse. Surprisingly, few researchers investigated such methods in the context of survey breakoff. Instead, it is common to see the adoption of complete case analysis in the past literature (Nissen, Donatello and Van Dusen, 2019). This practice assumes that missing data happens completely at random (Enders, 2010), which is hardly true in reality. Conducting analysis based on a wrong assumption about missing data is likely to result in biased estimates (Enders, 2010).

Given that breakoff is a special type of survey nonresponse, two common methods for mitigating the impact of survey nonresponse could offer fruitful avenues. These two methods are weighting and imputation. While the former weighs up/down the impact of under-/over-represented respondents, the latter uses answers from other respondents and questions to predict the missing data. Both methods have pros and cons.

Weighting is easy to calculate as it only requires users to build a statistical model to predict respondents' response propensity and then take its reciprocal. One problem with applying weighting to breakoff is that weighting is a unit-level statistic and its value remains constant for a given respondent, but breakoff happens at the item level and its propensity varies by questions. Also, breakoff weighting can only be given to complete respondents, which means breakoff cases are never included in the analysis even though they do provide answers to some of the questions.

In contrast to the weighting approach, where information from the breakoff cases (no matter whether they break off early or at a later stage) is never utilised in the analysis, imputation uses all valid answers from breakoff cases. This is because imputation is conducted at the question level. By including more data in the analysis, higher statistical power and, potentially, less biased survey estimates can be achieved. However, imputation also has its limitations, and one of the them is related to its complexity. For example, unlike the weighting approach where users build one model that explains the breakoff status, users of the multiple imputation might have to fit multiple models, each of which accommodates the types of missing variables (e.g., categorical, numeric). Also, if the variables in the data are related (e.g., the imputed number of open-ended questions should not exceed the number of total questions respondents saw in the survey), such a relationship should be included in the imputation algorithm. Moreover, before answering the substantive questions, users of the multiple imputation have to first create multiple imputed datasets, build the substantive models on them separately and then pool together the model coefficients (Enders, 2010). All these steps add complexity to the application of multiple imputation.

Both weighting and imputation have strengths and weaknesses when compensating for breakoff, but there is limited amount of current research that investigates which of them is more suitable for dealing with the breakoff bias and, more importantly, under what circumstances. Currently, instead of using any of these two methods to treat breakoff as an outcome separate from unit nonresponse, most surveys combine breakoff with unit nonresponse (e.g., Bailey *et al.*, 2017; CRONOS team, 2018). How such decisions may impact survey estimates is not documented. This paper fills this gap by comparing the effectiveness of different breakoff treatments in a simulation where the breakoff rates and causes are manipulated.

6.2 Background

6.2.1 Breakoff and missing data mechanisms

Breakoff is common in surveys. For example, Liu and Wronski (2018) reviewed 25,000 non-probability web surveys distributed among panellists of a commercial survey platform. They found an average breakoff rate of 13%. In another example, Revilla (2017) analysed 186 non-probability surveys conducted in another commercial web panel and reported a mean breakoff rate of 11.8%.

Breakoff leads to missing data, which reduces the available sample size and statistical power. In the worst situation, the survey estimates become biased if there are systematic breakoff patterns behind the missing data. For instance, in the study conducted by Steinbrecher, Roßmann and Blumenstiel (2015), voters who had not decided their favourite political candidate quit the survey more often, resulting in an underestimated proportion of undecided voters in the study.

Previous research on missing data has classified missing data based on three distinct mechanisms (Rubin, 1976; Enders, 2010). This classification can also be applied to breakoffs. The first type is Missing Completely At Random (MCAR). In this scenario, the occurrence of breakoff happens completely randomly, so the resultant data missingness is not related to any variable in the data (i.e. neither the outcome of interest nor other variables). The second mechanism is Missing At Random (MAR). In this case, breakoff happens because of some measured and observed variables in the data. For example, respondents who answer the survey on their mobile phones have a higher chance of breakoff (Chen *et al.*, 2022). As the information about the responding device is commonly recorded in web surveys, the breakoff event is now related to this measured variable. Finally, breakoff can happen due to a mechanism that is Missing Not At Random (MNAR). In this case, the occurrence of breakoff and missing data are directly related to the outcome of interest, and such information is not available to the researchers. The underestimated number of undecided voters in Steinbrecher et. al (2015) study previously reviewed is a real-world example of MNAR.

The review so far shows that different breakoff mechanisms lead to different missing data patterns. When dealing with such missing data, an appropriate method should be used to compensate for these different breakoff mechanisms.

6.2.2 Methods for breakoff compensation

Compensating for the missing data can be complex. As a result, researchers may simply discard cases that have missing data in any of the variables (i.e. listwise deletion) or in the variables of interest (pairwise deletion). For instance, Nissen, Donatello and Van Dusen (2019) reviewed 28 studies about physical education and reported 23 of them used complete case analysis. Simply removing any cases with missing data ignores the survey weights to compensate for the unit nonresponse. A more serious issue is that this approach assumes that the data are MCAR, thereby ignoring the possible mechanisms of MAR and MNAR. However, the occurrence of missing data due to a MCAR process is rare, and many past studies have documented that people with certain characteristics have higher chances of breaking off. Examples of such characteristics are age (those who are older are more likely to breakoff) and education (those with lower education are more likely to break off) (Peytchev, 2009; Blumenberg *et al.*, 2018). Therefore, MAR and MNAR are more plausible in practice.

Considering that breakoff is a special type of nonresponse, two post-collection techniques used to correct for the survey nonresponse can also be used for breakoff compensation. One is weighting, which is essentially a set of numeric values that weigh up/down certain respondents if they are under- or over-represented in the sample (Toepoel, 2015). One way to obtain the weighting is by taking the reciprocal of the survey response propensity scores generated by a statistical model that uses the binary response status as the dependent variable. Weighting has been implemented in many probability surveys not only because of the ease in its calculation but also due to its ability to reduce bias caused by survey nonresponse. For example, by comparing the prevalence of different diseases estimated from a probability survey in Sweden with the values from a national register database, Bonander *et al.* (2019) found most diseases were underestimated due to survey nonrespondents. However, the bias was reduced after nonresponse weighting was applied during the estimation.

The concept of weighting can be easily extended to breakoff using a two-step propensity model. In the first step, sample members' propensity to respond to the survey is estimated by a model, and the survey response weight is obtained by taking the reciprocal of the estimated response propensity. In the second step, a separate model is built to estimate the respondents' survey breakoff propensity, whose reciprocal becomes the breakoff weight. In the end, both weights are multiplied to form the combined weight. Theoretically, this combined weighting should account for not only the factors impacting both unit nonresponse and breakoff (e.g., gender found in Peytchev, 2011) but also those factors that are uniquely associated with breakoff (e.g., responding device found in Chen *et al.*, 2022). Therefore, both unit nonresponse and breakoff bias should be mitigated by this combined weighting approach.

Despite the ease of calculation, the weighting approach tends to lead to larger variations (i.e. less precision) in the analysis result (Biemer and Christ, 2008). This is due to large weights which can significantly increase the uncertainty of statistical estimates. This issue could be a concern for the combined weighting approach where two weights are multiplied, which has the risk of inflating the variation even further. Moreover, weighting is created at the unit level, so its value remains constant for a given respondent. In contrast, breakoff is an item-level event, and its propensity varies across questions even for the same respondent. It is unclear from the past literature whether breakoff propensity can be approximated well as a unit-level estimate.

Apart from weighting, multiple imputation is another popular approach that is used for compensating for missing data and can also be adapted to deal with breakoffs. According to Rubin (1987) and Curley *et al.* (2019), the first stage of the multiple imputation is to develop statistical models to explain the distribution of the missing variable given other variables in the dataset. Based on the developed model, imputed values are randomly drawn from the posterior distribution. This random drawing happens multiple times (typically five to ten) and results in multiple imputed datasets. At the second stage, users conduct their analysis separately on each of the imputed datasets. The results of the analysis are pooled in the third stage.

The multiple imputation approach makes use of answers not only from the complete respondents but also from those late breakoff cases (i.e. those who provide answers to some questions prior to their breakoff). Using all available information in the data can improve power and potentially minimise bias. Additionally, drawing imputed values multiple times from a distribution accounts for the uncertainty in the predicted values. This uncertainty is also propagated into the analysis stage, making the entire analysis more realistic (Enders, 2010).

However, the multiple imputation also comes with challenges. For example, given multiple imputed datasets, the data analysis becomes more complex. This is because researchers have to develop a model to impute the missing data before they can build the substantive model for their research questions. However, developing the imputation model is not easy because many decisions are involved (e.g., what imputation model should be used for each missing variable, what variables should be included in the imputation model).

Although weighting and imputation differ in how they compensate for breakoff bias, both of them require breakoff to be treated separately from survey nonresponse. However, many surveys in practice simply classify breakoffs as survey (non)response. For example, all breakoff cases were considered as survey respondents and given nonresponse weights in CRONOS, a probability web panel which will be described in more depth later in the Data Section of this study (CRONOS team, 2018). Meanwhile, designers of Next Steps Age 25 Survey, a longitudinal probability cohort study in England, adopt a slightly different definition of breakoffs. They selected a question block (which consists of questions of the same topic) and assigned breakoff cases to be nonrespondents if they broke off prior to the selected block (Bailey *et al.*, 2017). Despite the varying definitions of breakoff, none of the surveys treats breakoff as a separate survey outcome and specifically compensates for it.

It is unclear from the literature whether breakoffs should be treated separately. On the one hand, breakoff cases do respond to the survey request, so they are respondents by definition. Many previous studies documented that the majority of breakoffs happen at the beginning of the survey (Peytchev, 2009; Vehovar and Cehovin, 2014; Mittereder and West, 2021; Chen *et al.*, 2022). Some early breakoff cases might answer only a small number of questions at the beginning of the survey and then quit the survey, but those questions are usually not the key variables to the survey topic. Meanwhile, the others might simply log into the survey page, skim through the questionnaire and leave the survey website without answering any question. All those cases can therefore be classified as nonrespondents. On the other hand, although some factors simultaneously affect both nonresponse and breakoff (e.g., gender) (Peytchev, 2011), there are additional impacting factors for breakoff. According to previous research, some of these factors are question characteristics (e.g., matrix question) (Steinbrecher, Roßmann and Blumenstiel, 2015) and paradata (defined as data about the response process, e.g., question response time) (Mittereder and West, 2021). Treating breakoff the same as survey (non)response fails to consider those factors and thus risks breakoff bias. Therefore, the specific causes of breakoff-induced missing data should be included separately in the post-collection adjustment. The present study investigates these two opposing treatments of breakoff and answers the first research question:

RQ1. Does compensating for breakoffs separately in the post-collection adjustment help reduce the bias in survey estimates?

To answer this question, we will develop a simulation where the population benchmark is known. We then simulate the breakoff in the data and apply four different treatments of breakoff in the data analysis. While two treatments do not account for breakoff directly, the other two do. The survey estimates derived from these four treatments will be compared with the population benchmark to answer RQ1.

As no method is perfect, certain methods are likely to be effective in compensating for the breakoff in some scenarios but less useful in others. The scenarios can be defined by two varying elements: breakoff rate and missing data mechanism. As reviewed earlier, the breakoff rate can vary dramatically from survey to survey. This might have an influence on the effectiveness of the proposed compensation methods. For example, when filling in the missing data in a specific variable, multiple imputation uses other variables in the data as covariates in the model. A higher breakoff rate might cause more missing data in the covariates, and this can adversely impact the fit of the imputation model and subsequently damage the effectiveness of this compensation method.

Additionally, similar to survey nonresponse, breakoff can happen because of three missing data mechanisms, and different mechanisms might require different compensation strategies. Under the first mechanism, breakoffs happen completely at random (MCAR), so the breakoff bias is unlikely to be present and there might be no significant difference between the compensation methods in terms of bias reduction. Alternatively, the breakoff event can be associated with certain factors (e.g., age, responding device) and therefore can lead to some breakoff bias (MAR). To mitigate the bias under this mechanism, the compensation method should account for those factors in the data analysis. The final breakoff mechanism is related with the outcome of interest (MNAR). For example, people with past unpleasant experience (e.g., domestic violence) might consider the questions of those topics too disturbing and choose to quit the survey. Given that this cause of breakoff is not known or measured, the compensation method might need to incorporate the different missing data patterns directly in the analysis. Thus, the second research question is:

RQ2. How is the effectiveness of breakoff compensation affected by different breakoff rates and missing data mechanisms?

6.3 Data

Data used in this study comes from Wave 6 of the CROss-National Online Survey (CRONOS) panel. CRONOS is an online probability panel in Estonia, Slovenia and Great Britain. It was set up to test the feasibility of conducting an online, cross-national and probability survey (CRONOS team, 2018). The target population of CRONOS is individuals who live in private households in the three countries and are at least 18 years old. The sample members of CRONOS were recruited from the eighth wave of the European Social Survey, a face-to-face probability surveys in Europe. Each wave of CRONOS has different topics. In Wave 6, topics covered include attitudes towards income equality, society fairness and political efficacy. Wave 6 of CRONOS was conducted between January and February in 2018. In total, 1,812 people in the three countries responded to the survey in this wave, and the response rate is approximately 80%. Among those respondents, 110 broke off, resulting in a breakoff rate of 6% (CRONOS team, 2018). The data records the response status as well as the breakoff status of every sample member and will be referred to as *original data* throughout this study.

The first three columns under the *original data* in Table 6.1 provide a descriptive summary of the categorical variables used in this study. As can be seen, the proportion of sample units from the three countries is roughly evenly distributed while Great Britain has a slightly higher share in the sample composition. The majority of people voted in the last election and were born in the respective country where the survey was conducted. Additionally, most of them were above 34 years old, received education at the level of medium or above, lived in the urban area, worked in a paid job over the past seven days. Overall, respondents are pessimistic by stating that their influence on the politics is medium or low (93% of people). Also, as shown in Table 6.2, respondents reported a low level of satisfaction with their country's economy performance and trust in the parliament (both measures range from 0 to 10, and a higher score indicates more trust/satisfaction).

Table 6.1. Descriptive summary of categorical variables used in this study, separated by original, complete and synthetic data.

Variable	Original			Complete		Synthetic	
	N	% (incl. missing)	% (excl. missing)	N	%	N	%
TOTAL	**2260**			**1568**		**5000**	
Country							
Estonia	724	32	32	509	32	1623	32
Great Britain	860	38	38	558	36	1779	36
Slovenia	676	30	30	501	32	1599	32
Vote in the last election							
No	434	19	20	279	18	889	18
Yes	1693	75	80	1289	82	4111	82
Missing	133	6	-	-	-	-	-
Born in the country							
No	256	11	11	134	9	427	9
Yes	2004	89	89	1434	91	4573	91
Gender							
Male	983	44	44	667	43	2129	43
Female	1277	56	56	901	57	2871	57
Age							
15-24	193	9	9	87	6	266	5
25-34	426	19	19	253	16	811	16
35-54	860	38	38	618	39	1965	39
55-74	669	30	30	529	34	1698	34
75+	107	5	5	81	5	261	5
Missing	5	0	-	-	-	-	-
Married							
No	804	36	36	493	31	1566	31
Yes	1447	64	64	1075	69	3434	69
Missing	9	0	-	-	-	-	-
Urban							
No	788	35	35	564	36	1803	36
Yes	1472	65	65	1004	64	3197	64
Education							
Low	321	14	14	180	11	564	11
Medium	1165	52	52	817	52	2605	52
High	769	34	34	571	36	1831	37
Missing	5	0	-	-	-	-	-

Work in a paid job in the past 7 days							
No	807	36	36	561	36	1796	36
Yes	1453	64	64	1007	64	3204	64
Response at wave 6							
Nonresponse	421	19	19	0	0	-	-
Response	1812	80	81	1568	100	-	-
Missing	27	1	-	-	-	-	-
Breakoff at wave 6							
Breakoff	110	5	6	0	0	-	-
Complete	1702	75	94	1568	100	-	-
Missing	448	20	-	-	-	-	-
Device							
Computer	1075	48	59	964	61	3052	61
Phone	587	26	32	475	30	1532	31
Tablet	146	6	8	129	8	416	8
Missing	452	20	-	-	-	-	-
People can influence politics							
Low	1449	64	65	1015	65	3229	65
Medium	651	29	29	455	29	1451	29
High	146	6	6	98	6	320	6
Missing	14	1	-	-	-	-	-

Due to survey nonresponse and breakoff, two categorical variables in Table 6.1 suffer from a noticeable amount of the missing data. They are respondents' voting in the last election (6% of cases have missing voting outcome) and responding device (20%).

Among the numeric variables shown in Table 6.2, only the average response time in the original data has many missing data (20% of its data are missing). On average, participants took approximately 18 seconds to answer a survey question.

Table 6.2. Descriptive summary of continuous variables used in this study, separated by original, complete and synthetic data.

Variable	Data	Min	Max	Median	Mean	SD	N	Missing %
Average response time (second)	Original	2	273	16	18.16	12.85	2260	20
	Complete	3	102	16	17.63	8.60	1568	0
	Synthetic	3	101	16	17.76	8.72	5000	0
Satisfaction with the economy	Original	0	10	5	4.88	2.11	2260	1
	Complete	0	10	5	4.91	2.10	1568	0
	Synthetic	0	10	5	4.93	2.10	5000	0
Trust in parliament	Original	0	10	5	4.51	2.38	2260	1
	Complete	0	10	5	4.56	2.39	1568	0
	Synthetic	0	10	5	4.57	2.39	5000	0

6.4 Simulation

Respondents who have complete information in the sampling frame, paradata and three substantive questions in the survey are used as the base for the simulation. There are 1,568 of such respondents, and their descriptive summary can be found in Table 6.1 and Table 6.2 under the *complete data*. This dataset pools together the data from all three participating countries and is the base for the simulation in this study. One alternative is to conduct the simulation separately for each participating country and then combine the simulated datasets. However, countries are found to be unrelated to the breakoff outcome. Also, because the *original data* have only 110 breakoff cases across the three participating countries, splitting this data into three subsets will greatly reduce the breakoff size in each subset and damage the effectiveness of the simulation. Therefore, the *original data* where three participating countries' records are pooled together will be used in the simulation in this study.

In total, the simulation has three stages (described below). All variables involved in the simulation are listed in Table 6.3 where a checked mark indicates the specific stage of the simulation at which these variables are used.

Table 6.3. Variables used in the simulation and their usage at different stages (Stage 1: create synthetic population; Stage 2: simulate survey nonresponse; Stage 3.1: simulate MCAR; Stage 3.2 simulate MAR; Stage 3.3 simulate MNAR).

		Usage in Simulation Stage			
Group	**Variable**	**1**	**2**	**3.2**	**3.3**
Sampling frame information	Vote in last election, Native citizen, Gender, Age, Marital status, Living in urban areas, Education, Work in a paid job in the past seven days, Survey participation country	✓	✓	✓	✓
Paradata	Device, Average question response time	✓		✓	✓
Substantive	Trust in parliament, Perceived political influence, Satisfaction towards economy performance	✓			✓

Stage 1. Create synthetic population and obtain benchmark values

The first stage is to increase the size of complete cases of the variables used in this study from 1,568 to 5,000 by creating a synthetic dataset. This is necessary because the number of complete respondents in the original survey data is too small for simulating survey nonresponse and breakoffs as well as deriving reliable survey estimates at the later stage. The R package *synthpop* is used, which looks at the patterns in the complete data and synthesises new cases based on the pattern (Nowok, Raab and Dibben, 2016).

Using only the complete cases might make the resultant synthetic population differ from the true general population. However, this decision is still considered appropriate because the main aim of this study is to test whether the breakoff compensation methods can mitigate the bias caused by breakoffs. Generalising conclusions to the true general population will require the other post-survey adjustments (e.g., compensation for coverage bias, weight trimming), which are additional steps beyond the focus of this study. To avoid confusion between the simulated population in this study and the true general population, the population created in this study will be referred to as *synthetic population*, and its population parameter will be called the benchmark value. Therefore, the 5,000 cases are the *synthetic population*. The distribution of all the variables before and after the synthesis can be found in Table 6.1 and Table 6.2 under the *complete* and *synthetic data*.

Using the synthetic population dataset, we can calculate the benchmark value for the outcome of interest in this study. That is, people's stated trust in the parliament of their country. This variable is chosen because it is an important measure of government's ability to perform its duties and is often used by researchers (e.g., van der Meer, 2010; Torcal and Christmann, 2021).

We calculate two statistics of interest for the stated trust in the parliament. The first one is its mean value. The second one refers to the coefficients of two explanatory variables of a linear regression model on people's trust in parliament: people's perceived influence on the country's politics and their satisfaction towards country's economy performance (sampling frame variables are used as control variables in the model). The benchmark values for these two sets of statistics will be calculated using the *synthetic population* dataset as survey nonrespondents and breakoff cases do not exist in this dataset at this stage.

Stage 2. Simulate survey nonrespondents

A logistic regression model using the survey response status as the dependent variable (1 = survey response, otherwise 0) is built using the *original data* (i.e. the data before the synthetic data is generated) to estimate the regression coefficients for the explanatory variables of survey response. The explanatory variables in this model are the sampling frame information such as their age, gender, voted or not in the last election (See Table 6.3). These variables are chosen because the CRONOS team used them to compensate for survey nonresponse (CRONOS team, 2018).

To obtain the target total of nonrespondents to be drawn later, we multiply the nonresponse rate in the original survey (i.e. 20%) with the size of the synthetic population dataset (i.e. 5,000), which results in 1,000. To simulate this number of survey nonrespondents, we first apply the above logistic model to the *synthetic population* data to predict the survey response propensity for every person. Following this, we define strata based on deciles of the predicted response propensities. Next, we calculate the mean of the response propensities in each stratum. We then randomly draw cases from each of the ten strata with probabilities proportionate to the mean response propensities. As such, people in the first quantile have the highest chance to be nonrespondents while those in the tenth quantile have the lowest chance. The simulated nonrespondents' answers to the three substantive questions and paradata in the simulated data are deleted, but their sampling frame information is kept in the data.

By artificially relating the survey (non)response with only the observed variables (i.e. sampling frame information), the survey nonresponse is simulated under the Missing At Random mechanism. Making this assumption has implications for the breakoff compensation. As will be described later, when simulating breakoffs, apart from using some variables that uniquely affect breakoff, the sampling frame information is also used. In this case, survey nonresponse and breakoff share some impacting factors, and compensating for the survey nonresponse bias will also compensate for the breakoff bias to some extent. This is likely to confound the performance of some breakoff compensation methods investigated in this study. However, because of the following reasons, we still conduct the simulation of survey nonresponse using the Missing At Random assumption (rather than other two missing mechanisms). Firstly, previous studies have noted that Missing Completely At Random rarely happens, so using an unlikely scenario for survey nonresponse will damage the application of our simulation. Alternatively, using the Missing Not At Random assumption for survey nonresponse will mean that we cannot fully accounted for the influence of some of its impacting factors on the survey estimates. Consequently, the nonresponse bias that is not compensated for will be mixed with the breakoff bias, making it difficult to evaluate the true effectiveness of the breakoff compensation methods. Secondly, past studies have already documented that some factors can impact survey nonresponse and breakoff simultaneously, such as gender (Peytchev, 2011). Therefore, to allow the simulation to be grounded in a real-world situation and to measure the true effectiveness of the breakoff methods, we decide to simulate the survey nonresponse using the MAR assumption.

Stage 3. Simulate breakoff cases

We simulate breakoff cases while varying two elements: the breakoff rate and breakoff mechanism. More specifically, we simulate four different breakoff rates, ranging from 5% to 20% at an increment of 5%.[4] The three breakoff mechanisms are MCAR, MAR and MNAR.

[4] We restrict the maximum of the breakoff rate in this simulation study to be no more than 20%. This is because any breakoff rate above 20% is uncommon according to the literature reviewed earlier.

These three mechanisms are independently simulated, and all of them are conducted only among respondents in the simulated data from Stage 2.

Stage 3.1. Missing completely at random

A proportion (as specified by the breakoff rate) of respondents in the data from Stage 2 are randomly assigned to be breakoff cases. Under this mechanism, no variable is used in the breakoff simulation, so the occurrence of breakoff does not depend on any variables in the data.

Stage 3.2. Missing at random

To simulate breakoff under the MAR situation, the *original data* is used to develop a logistic regression model where the outcome variable is the binary survey completion status and the explanatory variables are sampling frame information and paradata (See Table 6.3). This model is then applied to the data from Stage 2 to predict the survey completion propensity for each respondent there. In this way, we ensure that breakoff event is related to sampling frame information and paradata. Based on the predicted survey completion propensities, respondents are assigned to one of the three quantiles, representing low/medium/high-breakoff-propensity groups, respectively.

We use a disproportionate approach to simulate breakoffs in MAR. To be more specific, we adopt a ratio of 1:2:10 when drawing breakoffs from the low, medium and high breakoff-propensity groups. As such, we ensure that people in the high-breakoff-propensity group have the highest chance to break off (i.e. with a probability of 10/13) while the other two groups have a much lower chance (2/13 and 1/13). We choose this ratio in order to test the extreme situation where the majority of breakoffs happen in the most-likely-survey-breakoff group. By multiplying the simulated breakoff rate (5%, 10%, 15% etc.) and the respondent size from Stage 2 (approximately 4,000), we can get the target breakoff size. We then randomly draw this size of breakoff cases from the three quantile groups respectively.

Stage 3.3. Missing not at random

In MNAR, apart from assigning respondents to one of the low/medium/high-breakoff-propensity groups as in MAR, they are also assigned to one of other three quantiles. These three new quantiles represent their low/medium/high trust in the parliament (i.e. our outcome of interest). The creation of the two quantile groups is independent from each other.

The cross-tabulation between the two quantile groups leads to a 3x3 table (See Table 6.4). We keep the 1:2:10 breakoff ratio for the quantile group about the survey breakoff propensity. For the quantile group about the parliament trust, we use a breakoff ratio of 5:2:1 for the low/medium/high level of parliament trust. As such, we can test the extreme situation where people who have low trust in their parliament break off more often. Also, we assume that, compared to the outcome of interest, the survey breakoff propensity has a larger influence on the survey breakoff likelihood, so we choose five (rather than ten) as the value for low parliament trust group.

Multiplying the ratios of the two quantile groups leads to the final ratio for simulating breakoffs in different quantile groups under the MNAR mechanism. For example, as shown in Table 6.4, if respondents score high in the survey breakoff propensity but low in the parliament trust, their chance of breaking off is 50 times higher than those who have a low propensity to break off but a high level of parliament trust (i.e. a 1/104 breakoff chance).

Table 6.4. Breakoff probabilities of different quantile groups under the MNAR mechanism.

	Trust in parliament		
	Low	**Medium**	**High**
Survey breakoff propensity			
Low	5/104	2/104	1/104
Medium	10/104	4/104	2/104
High	50/104	20/104	10/104

We calculate the target breakoff total to be drawn as before and then randomly draw the breakoff cases from the nine quantiles disproportionately using the probabilities in Table 6.4. By including the trust in parliament when simulating breakoffs but deliberately not accounting for it when conducting analysis later, we create a Missing Not At Random situation.

In the end, we create 12 breakoff datasets (4 breakoff rates × 3 breakoff mechanisms). Note that in comparison to Stage 2 where the dataset has only nonresponse but no breakoff, these 12 datasets now have both nonresponse and breakoff. For all breakoff cases, their answers to the three substantive questions are discarded, but their sampling frame information and paradata are kept in the data. To account for the variation in the simulation, we repeat each of the three simulation stages 200 times, so we create 2,400 datasets to be analysed (4 breakoff rates × 3 breakoff mechanisms × 200 repetitions). All simulation and data analysis in this study are conducted in R 4.1.2 (R Core Team, 2021).

6.5 Analysis plan

As mentioned earlier, we calculate two statistics of interest (i.e. mean parliament trust and regression coefficients) for each of the 12 breakoff scenarios. We conduct the calculation with four compensation methods for breakoff.

1. **Ignore breakoff (IG):** The first method assumes that breakoffs do not happen and simply discards them in the analysis. As a result, only the nonresponse weights are incorporated into the analysis. To calculate the nonresponse weights, a logistic regression is fitted on the simulated data from Stage 2 (i.e. simulated population with nonrespondents only). This model uses the binary survey response status as the dependent variable (1 = survey response) and sampling frame information as the explanatory variables (see Table 6.3). The fitted model then outputs the response propensity, and its reciprocal becomes the nonresponse weights. The process of deriving the nonresponse weights remains the same in the third and fourth method described below, so it is no longer explained later for brevity. Note that no calibration is used in this study. Also, although everyone in the data will be assigned a nonresponse weight, the data analysis is based only on the survey respondents. The first breakoff compensation method helps us investigate whether completely ignoring breakoff influences the survey estimates and if so to what extent.

2. **Treat breakoff as nonresponse (NR):** The second method recognises the occurrence of breakoff but does not treat it as a special survey outcome. Instead, it classifies all breakoff as survey nonresponse. As now the number of nonrespondents increases, a logistic regression specified in the same way as in the first method is fitted again but

this time on this new dataset. The response propensities and survey nonresponse weights are updated accordingly. Our two types of statistics of interest are calculated based only on respondents who have the updated survey nonresponse weights. As reviewed earlier, this method is currently being implemented by some surveys. We are interested in how this method performs in terms of mitigating the bias caused by the breakoff.

3. **Combined weighting (CW):** Unlike the previous two methods, the combined weighting approach not only recognises the presence of breakoff but also compensates for it separately. It considers breakoff as a survey outcome conditional on the survey response and calculates weights dedicated for breakoff. Calculating such weights requires fitting a separate logistic regression model. The dependent variable in the model is the breakoff status (1 = survey completion and 0 = survey breakoff), and the explanatory variables are sampling frame information and paradata listed in Table 6.3. Following this, respondents' propensity to complete the survey is predicted by the newly fitted logistic model, and its reciprocal becomes the breakoff weights. The process of deriving the breakoff weights is the same in MCAR, MAR and MNAR. Only complete respondents have breakoff weights. Finally, the combined weighting is obtained by multiplying the breakoff weights and nonresponse weights and then used in the calculation of our statistics of interest. This combined weighting approach will shed light on whether using the unit-level breakoff weighting can remedy the bias caused by the item-level breakoff event.

4. **Multiple imputation (MI)**: This method fills in the missing data in breakoff respondents' answers to the three substantive questions using values from complete respondents who share similar characteristics with them. The predictive mean matching algorithm is used in the imputation. According to Vink *et al.*(2014), this algorithm first fits a model with all other variables as independent variables (i.e. sampling frame information + paradata + substantive variables + binary survey breakoff status). It then generates predicted values for the variable under imputation for both the complete respondents and missing cases. Following this, the algorithm finds five complete respondents that have the closest prediction to the case with the missing data. One of these five respondents is then randomly chosen as the donor, and

the missing cases' predicted value is replaced by the observed value from the donor. Imputing missing data with the observed values avoids invalid or extreme imputed values. To account for the uncertainty and variability in the predicted values, we create 10 imputed datasets for each of the 12 simulated breakoff scenarios. The process of replacing missing data operates iteratively in each imputed dataset and the first few iterations usually lead to volatile values. To help the convergence of the imputed values, for each of the 10 imputed datasets, we allow the algorithm to replace the missing values 50 times before finally filling in the missing data with the value. Once the imputation is completed, each of the 10 imputed datasets will produce its own estimate for our statistics of interest, which is then pooled together using the combining rule of Rubin (1987). They are then weighted by the existing nonresponse weights, which is calculated in point (1) above, to create the final survey estimates. The multiple imputation is conducted in R using the *mice* package (van Buuren and Groothuis-Oudshoorn, 2011).

In the end, we calculate our statistics of interest under 48 different situations (i.e. 12 simulated breakoff scenarios × 4 breakoff compensation methods). We then compare them with the benchmark value obtained at Stage 1 using the relative absolute bias (RAB).

$$RAB = \sum_{i=1}^{200} \left(\frac{|x_i - X|}{|X|} \right) \times \frac{1}{200} \times 100\%$$

The equation above shows the calculation of RAB for our statistics of interest where x_i (i = 1, …, 200) represents the survey estimates derived by applying the four treatments of breakoff to each of the 200 simulated breakoff datasets. The benchmark value (denoted as *X*) is calculated by taking the average of the corresponding statistic of interest from the 200 *synthetic population* data.

By comparing the RAB among the four treatments of breakoff, we will know how they perform under different breakoff rates and mechanisms in terms of mitigating the breakoff bias. This comparison also allows us to investigate how different types of survey estimates (i.e. univariate estimates and model coefficients) are affected by the breakoff rates and mechanisms.

6.6 Result

Figures below display the Relative Absolute Bias for the two statistics of interest in this study: (1) the mean of parliament trust, and (2) the model coefficients regarding the impact of people's perceived political influence (a categorical variable with three levels) and satisfaction with their country's economy (a continuous variable) on their parliament trust. All figures are separated by breakoff rates (right, row) and breakoff mechanisms (column). Each point in the figures represents the RAB value resulted from using a specific method for dealing with breakoffs. The exact RAB values are provided next to the corresponding points. Due to rounding, some points located differently on the horizontal scale might have the same RAB values next to them.

For the detailed distribution of the two statistics of interest in this study, see the boxplots in Appendix C. The Root Mean Square Error (another metric commonly used to evaluate the simulation) is also calculated but not reported here because it leads to the same conclusion as RAB. For details about RMSE, see Appendix D.

6.6.1 Mean of parliament trust

As can be seen in Figure 6.1, under the MCAR situation, there is no difference between RAB of the four methods when estimating the mean parliament trust. In MAR, the RAB is slightly different between the four methods, but the difference is so small that all four methods are still considered to be equally capable of correcting for the bias in the mean of parliament trust.

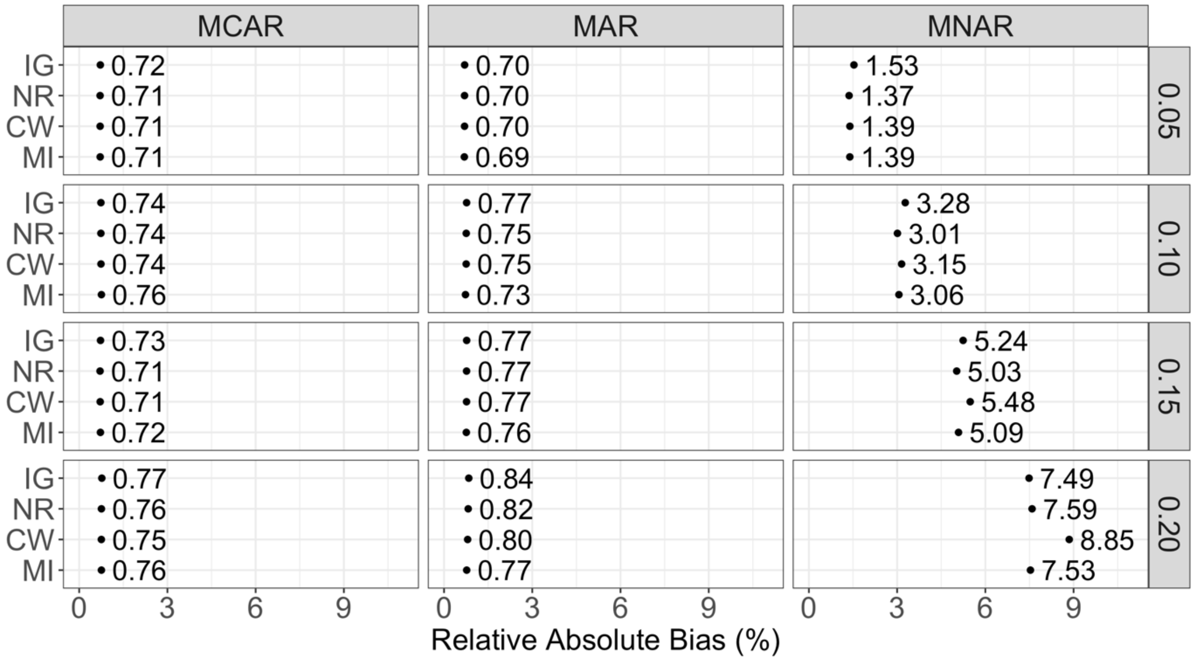


Figure 6.1. Relative Absolute Bias of the mean parliament trust, estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

Under the MNAR, all four compensation methods clearly produce a biased estimate of the mean parliament trust. Indeed, as the breakoff rate increases, the RAB climbs from approximately 1% to 8%. It is noteworthy that the combined weighting approach performs similarly to other methods when the breakoff rate is 5% and 10%. Nonetheless, when the breakoff rate is 15% and above, this approach starts to exacerbate the bias in the estimate. In fact, it becomes the worst performing method in the 20% breakoff rate scenario. Another finding is that methods which do not treat breakoff specifically (i.e. ignoring breakoff or treating breakoff as nonresponse) perform equally well compared to multiple imputation in terms of RAB.

6.6.2 Model coefficients

Figure 6.2 and Figure 6.3 show the regression coefficient for the medium and high level of the perceived political influence (reference: low) on parliament trust, respectively. As before, four methods perform almost equally well under MCAR.

Under MAR, for the coefficient of medium political influence, the RAB is still similar across the four compensation methods. However, for high political influence, a split in the compensation performance emerges. Indeed, when the breakoff rate is 20%, combining the breakoff and nonresponse weighting leads to more biased estimated regression coefficients while multiple imputation produces the least biased result. The RAB for the two methods that do not correct for breakoff directly falls in between.

For the MNAR mechanism, the findings are mixed. For the medium level of political influence, ignoring breakoff generates the lowest RAB value, and the worst performing method is the combined weighting approach. However, for the high level of political influence, ignoring breakoff produces the worst RAB among the four methods while multiple imputation has the smallest RAB. For both medium and high level of people's perceived political influence, treating breakoff as nonresponse (the method currently used by many survey agencies for dealing with breakoff) is never the first in the ranking of the best performing method.

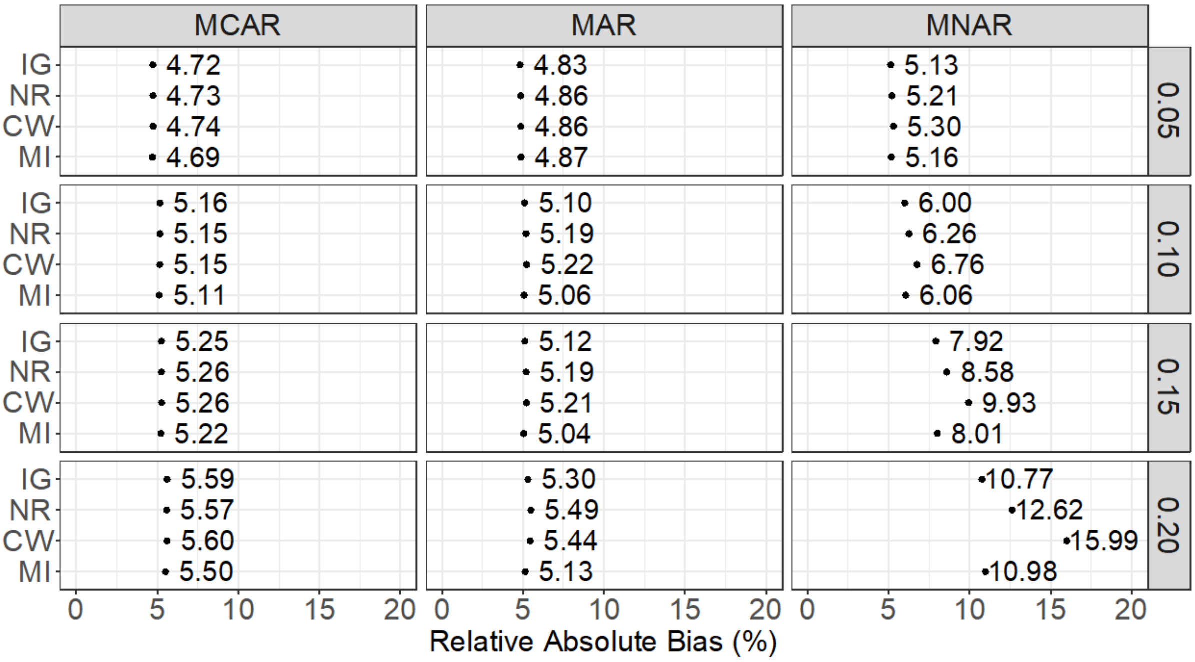


Figure 6.2. Relative Absolute Bias of the model coefficient corresponding to the perceived political influence (medium vs. low), estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

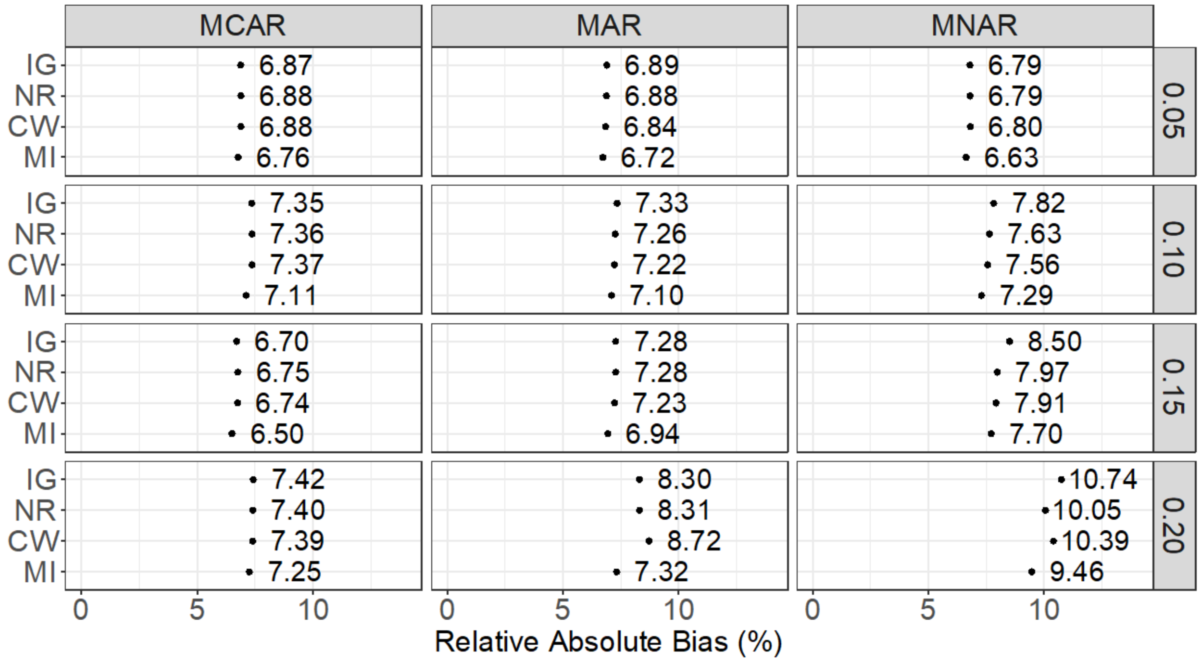


Figure 6.3. Relative Absolute Bias of the model coefficient corresponding to the perceived political influence (high vs. low), estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

When it comes to the RAB for the regression coefficient of people's satisfaction with their country's economy performance on parliament trust (See Figure 6.4), there is not much difference in the RAB among the four methods. This is true given that the difference in RAB is less than 1% among the four compensation methods for any combination of missing data mechanisms and breakoff rates. It means all four methods can perform equally well when mitigating the bias in the model coefficient of people's economy satisfaction on parliament trust.

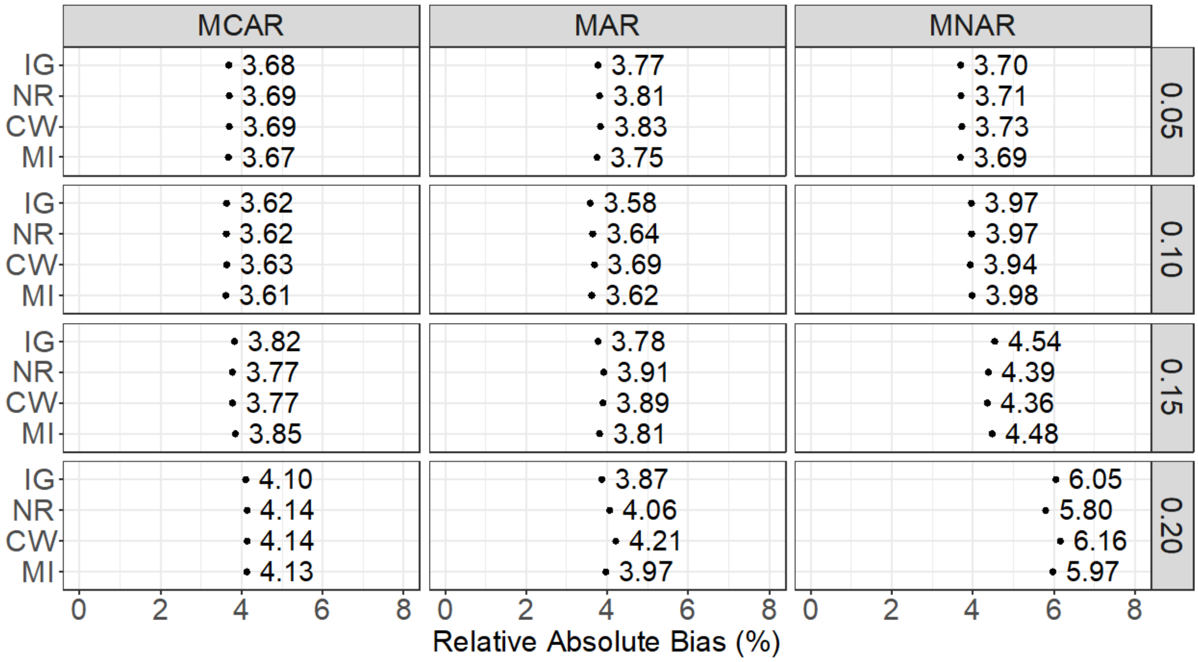


Figure 6.4. Relative Absolute Bias of the model coefficient corresponding to the satisfaction towards country's economy performance, estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

Overall, looking at the RAB of the four compensation methods across the univariate mean and regression coefficients, the multiple imputation is found to give a good and consistent performance when compensating for breakoff as it produces either the lowest or the second-lowest RAB depending on the variables. One possible reason is related to how this method imputes the missing data. As multiple imputation operates at the question level, it can learn the complex relationship between missing data and observed variables at the question level. Taking into account this subtle relationship might improve the compensation. In contrast, the combined weighting approach can only consider the same relationship at an aggregate level (i.e. respondent level) and therefore fails to capture the subtlety in the relationship and becomes the worst-performing method in this study.

For the compensation method that is currently used by many survey agencies (i.e. breakoff is considered as unit nonresponse and compensated for via nonresponse weighting), it is rarely the best-performing method in this study. This means that not considering the factors that uniquely impact the breakoff fails to correct for the breakoff bias. As for the approach that ignores the breakoff (i.e. IG), it sometimes surprisingly generates a better or similar performance compared to other three methods that acknowledge the existence of breakoffs.

One possible explanation for this finding is that some survey variables might not have a large amount of breakoff bias to be compensated for. As the impact of breakoff on different variables is unlikely to be universal, some variables might suffer from little breakoff bias. When the amount of breakoff bias is small (especially when the breakoff rate is low), the breakoff cases can be discarded without greatly damaging the estimates. However, in such a situation, for the methods that either update the nonresponse weights by treating breakoff cases as nonrespondents (i.e. NR) or create weights specifically for breakoff (CW), the resultant weights might be too large for a small amount of breakoff bias. As a result, both methods overshoot the benchmark values. Given the good performance of multiple imputation in such a situation as well as its ability to keep the maximum amount of information in the data (as opposed to discarding the answers from the late breakoff cases), this method should be recommended for the breakoff compensation.

Moreover, comparing the RAB of the univariate mean and regression coefficients gives an insight into how different types of estimates are affected by the breakoff. To be specific, comparing the RAB of the mean parliament trust (i.e. Figure 6.1) to that of the three model coefficients (i.e. Figure 6.2, Figure 6.3, and Figure 6.4), it can be found that the former is smaller than the latter for any given combination of the breakoff rate and missing data mechanism. Indeed, the RAB of the mean parliament trust is between 0.7% and 8%, but that of the model coefficients ranges from 3% to 16%. All this implies that the breakoff is more likely to affect the multivariate model coefficient than the univariate estimate.

Comparing the RAB between the univariate mean and regression coefficients also shows how sensitive different types of estimates are to the change in the missing data mechanism. More specifically, the RAB of the mean parliament trust is close to 0.7% under the MCAR and MAR scenarios, but this value can increase by as much as 7% when MNAR happens. In contrast, the same shift in the missing data mechanism only leads to a moderate increase (less than 3% most of the time) in the RAB of the model coefficients. An exception is the model coefficient corresponding to the perceived medium political influence. When the breakoff rate is 20% and the missing data mechanism changes from MCAR/MAR to MNAR, the RAB of this coefficient increases by as much as 10%. Nonetheless, for other three breakoff rates (5% to 15%), when the missing data mechanism changes, the change in its RAB is relatively moderate. Overall, we can conclude that the univariate estimate of the outcome variable (as opposed to the model coefficients of the explanatory variables) is more sensitive to the shift from the MCAR/MAR to MNAR.

6.7 Discussion

Many researchers have proposed different design optimisations or real-time interventions to discourage breakoff, but few have investigated the impact of post-survey adjustments on survey estimates. This study fills in the gap by comparing four methods that compensate for breakoff, with a focus on testing whether methods that treat breakoff as a special survey outcome can bring extra benefits in terms of correcting for the bias (RQ 1). The robustness of all four methods is also investigated under different breakoff rates and missing data mechanisms (RQ 2).

The answer to RQ 1 is mixed and needs further research. To be specific, compared to the two methods that do not treat breakoff directly (i.e. ignoring breakoff or treating breakoff as unit nonresponse), calculating weights for breakoff and combining it with the nonresponse weights does not reduce breakoff bias. In fact, this combined weighting approach can exacerbate the bias for some variables. This finding indicates that breakoff should not be treated separately from unit nonresponse. However, when comparing multiple imputation to the two methods that do not treat breakoff directly, the former helps reduce the breakoff bias to some extent. This finding implies that breakoff should be compensated for separately. Nonetheless, the statistical model used in the multiple imputation (i.e. predictive mean matching) differs from that in the weighting approach (i.e. logistic regression). In this case, it is unclear whether the superior performance of multiple imputation is due to either the model used or the separate treatment of breakoff.

In terms of the robustness of the four compensations methods (RQ 2), it is found that their performance is affected by the breakoff mechanism more than the breakoff rate. To be specific, as the breakoff rate increases within the same breakoff mechanism, the performance ranking of the four methods does not change (except for the combined weighting approach in the univariate mean estimation). When the breakoff rate remains the same but the breakoff mechanism changes, the best compensation method varies. Therefore, the discussion on the robustness of the four methods employed in this study will be based on the missing data mechanisms.

When the data are MCAR, no difference is found in survey estimates derived from the four compensation methods. This is true for the estimation of the univariate mean as well as for the model coefficients in the multivariate analysis. Nonetheless, the finding is less consistent under the MAR mechanism. For some variables, all four methods perform equally well, but for others multiple imputation gives the smallest relative absolute bias value while the combined weighting is the worst performing method. Under the MNAR mechanism, the combined weighting approach tends to produce the worst performance, and the performance of the remaining three methods varies depending on the statistic of interest.

Overall, the answer to RQ 2 is as follows. For MCAR, there is no dramatic difference in the performance among the methods. For MAR, multiple imputation helps mitigate the breakoff bias but only slightly better compared to other three methods. Meanwhile, correcting for breakoff using the breakoff weights usually leads to more bias. In the context of MNAR, none of the four methods in this study solves the breakoff bias.

Based on the findings, there are two practical implications for the survey designers. Firstly, more attention should be paid to the cause of breakoff. This study has shown that the occurrence of breakoffs can affect the survey estimates in both univariate and multivariate analysis, and more importantly, different causes of breakoff impact different types of estimates disproportionately. Currently, many survey agencies either do not record the breakoff incident or simply combine the breakoff cases with partial interviews or unit nonresponse. In this case, it is impossible to investigate whether breakoff bias exists and its extent.

Secondly, researchers should use multiple imputation to compensate for the breakoff-induced missing data. In the simulation, the multiple imputation shows a good and consistent performance (i.e. it is either the best or the second-best performing method in the simulation depending on the variables). One might argue against the recommended multiple imputation as, for some variables, not directly correcting for breakoff generates similar relative absolute bias as the multiple imputation. However, even in this case multiple imputation can include the substantive answers from late breakoff cases in the analysis, which should improve the statistical power. It should also be highlighted that the missing data pattern in this simulation is monotonic (i.e. breakoff cases will always have all their substantive answers removed no matter at which substantive questions they break off). When there are more missing data patterns in the variables, the ability to use more information in the analysis is expected to make multiple imputation perform better in reducing the bias.

The present study has some limitations that future studies can address. To begin with, there are only two factors that uniquely impact breakoff (i.e. responding device, and average question response time) in the simulation. Researchers can conduct a simulation where breakoff has more of its own impacting factors (e.g., mouse back clicks, the number of survey logins) to investigate whether the combined weighting approach becomes effective in reducing the bias under such a scenario. Secondly, as no methods in this study can solve the breakoff bias under the MNAR mechanism, future research should investigate other solutions such as pattern mixture models (Hedeker and Gibbons, 2006). Moreover, once a person is chosen in our study to be the breakoff case, all their substantive answers are removed in the simulation. However, people can break off at different questions in practice, meaning that for some breakoff cases we can have a record of their answers to some substantive questions (as opposed to seeing all the answers missing). Having such diverse missing data patterns will allow researchers to explore how the effectiveness of breakoff compensation methods is affected when the variables suffer from different proportions of missing data. Furthermore, we cannot find variables that (1) are available to both complete and breakoff respondents and (2) strongly correlate with both the breakoff propensity and the outcome of interest. This might limit the effectiveness of the two methods that account for breakoff directly. Future research can replicate this study with variables that satisfy the above two criteria to see whether the findings here still hold.

Despite the limitations, our study provides empirical evidence on the effectiveness of commonly used practice where the breakoff is considered as unit nonresponse and compensated for via nonresponse weighting. Neither theoretical nor empirical justification for this breakoff treatment was documented until now. Also, our study demonstrates that the breakoff, as an item-level event, should be accounted for by the item-level technique such as multiple imputation. Using a respondent-level breakoff weighting could exacerbate the breakoff bias issue.

Chapter 7 Conclusion

This thesis focuses on tackling the web survey breakoff by understanding, predicting and mitigating this event at three stages of the survey: before, during and after the survey data collection. Findings in this thesis will contribute to the theoretical development in the field of survey breakoff research. Also, this thesis can provide guidance for the survey designers to tackle the breakoff issue in practice.

7.1 Theoretical contributions

Chapter 4 demonstrates the importance of accounting for the breakoff timing when studying survey breakoff. Currently, when investigating whether some survey designs lead to more breakoffs, most past studies ignored the breakoff timing. When researchers found that the different designs under investigation led to the same amount of breakoff, they usually concluded that the designs did not affect the survey breakoff (e.g, Healey, 2007; Hohne, Schlosser and Krebs, 2017). However, this ignores the fact that some designs might be able to postpone the breakoff event, and such designs will be preferred as researchers want the respondents to answer more questions prior to the breakoff event. By taking into account the breakoff timing, Chapter 4 shows that respondents in the interleafed format quit the survey earlier in comparison to the grouped format. This insight is new as the breakoff timing was not evaluated in the past relevant studies (Kreuter *et al.*, 2011; Eckman *et al.*, 2014). This new insight also implies that respondents' breakoff decision is likely linked to the moment when they learn the extra burden caused by the filter questions. All of this contributes to the understanding of the process leading to the breakoff.

Additionally, Chapter 4 is the first published study that uses an experimental design to causally investigate the effect of question topic on breakoff. Surveys normally have multiple topics arranged in a specific order, and the response burden accumulates during the survey answering process. Previous research has shown that both the order of the question topics and the duration respondents have spent on the questionnaire are associated with breakoff (Hoerger, 2010; Teclaw, Price and Osatuke, 2012). Thus, studying the real impact of question topics on breakoff requires the separation of these two confounders. However, previous research failed to do this as the ordering of questions in those studies was always fixed. Using a randomised experiment and the Cox survival model, Chapter 4 shows that the introduction statement (i.e. statement that introduces the upcoming topics but does not ask any question of substantive topics) and the topic about respondents' non-health insurance (e.g., vehicle and home insurance) were associated with more breakoffs, in comparison to respondents' clothing purchase and utilities payment. The experimental design in Chapter 4 allows us to conclude that these two topics are genuinely associated with more breakoffs. Researchers can proceed to investigate the underlying reason in the future.

Chapter 5 contributes to the idea of intervening in real time to minimise survey breakoffs. Firstly, it contributes to the development of the statistical model that can accurately predict the question-level breakoff. To be specific, Chapter 5 finds that gradient boosting outperformed the traditional Cox survival model in breakoff prediction as it achieved a good balance between the true positive and false positive rate. In addition to this, gradient boosting has fewer model assumptions than the Cox survival model. It means that researchers need not worry about any possible violation to the proportional hazard assumption in the Cox survival model. Moreover, the tree structure of gradient boosting can implicitly accommodate the interaction effects among predictors. All of this indicates that gradient boosting is preferred to the traditional Cox survival model for predicting breakoff in real time. More importantly, this finding can contribute to a broader discussion: whether the machine learning models can perform well in predicting breakoff when there is a clustered structure in the data (i.e. questions are clustered within respondents). As machine learning models are not designed to handle the specific challenges of breakoff data (e.g., clustering, censoring), many researchers are currently using the Cox survival model to explain the breakoff (Peytchev, 2009; Hochheimer *et al.*, 2016; Mittereder and West, 2021). To the knowledge of the author, Chapter 5 is the first publication that compared both survival models and machine learning models in the context of breakoff prediction. In the end, it was surprising to see that taking extra account of the special structure in the breakoff data did not lead to an improved performance in the prediction.

The second contribution from Chapter 5 is the clarification on the way to code time-varying predictors to maximise the model performance in predicting breakoff. Some of researchers accumulated its value from the beginning of the survey (e.g., Peytchev, 2009) while others did not carry out any manipulation, using them concurrently (e.g., Vehovar and Cehovin, 2014). While the assumption behind the cumulative coding is that the response burden accumulated since the start of the survey drives the breakoff event, the concurrent coding assumes the burden experienced in the moment matters the most. Chapter 5 is the first study, to our knowledge, that systematically investigates these different coding schemes and the associated assumptions, particularly in the context of prediction performance. The concurrent coding of time-varying variables was found to be more predictive of the breakoff event than the cumulative coding. This demonstrates that the current response burden has more influence on respondents' breakoff decision than past cumulative burden. This fits into one of the findings in Chapter 4 (respondents might break off in the moment they realise the burden).

Chapter 6 contributes to the literature on post-survey adjustments from two perspectives. To begin with, it is one of the few studies that provides empirical evidence on how the statistical estimates are affected when breakoff is treated as a survey outcome separate from unit nonresponse. Currently, breakoff is commonly classified as unit nonresponse and compensated for using nonresponse weighting (e.g., Bailey *et al.*, 2017; CRONOS team, 2018), but this approach was never the best performing method in our simulation. This finding casts doubt on the current practice for breakoff compensation. Not only is this practice being used without any justification in the documentation, but also it risks not fully accounting for the missing data caused specifically by breakoffs. As a result, the survey estimate may be biased. Currently, the discussion on whether breakoff should be treated as a unique survey outcome is sparse, and research on post-survey breakoff treatment is still in its early stages. Nonetheless, it is believed that Chapter 6 will provide a good starting point for such research in the future.

Furthermore, Chapter 6 is the first study, to the knowledge of the author, to compare weighting and imputation to correct for breakoff. Previous research only applied weighting to adjust for breakoff (Steinbrecher, Roßmann and Blumenstiel, 2015). Unlike weighting which is a respondent-level technique, multiple imputation operates at the question level and is expected to handle the breakoff well because breakoff is a question-level event. This speculation was not tested in the published literature until now. It is found in Chapter 6 that multiple imputation produced a lower bias in the estimates than the weighting approach. Additionally, imputation has the benefits of keeping more information in the data than weighting (where the breakoff cases are excluded even though they provide answers to some questions). All of this indicates that breakoff should be dealt with using multiple imputation.

7.2 Practical recommendations

In addition to the contribution to the academic literature, this thesis also helps inform practical decisions that can be taken to tackle the breakoff throughout the entire cycle of survey data collection. Firstly, knowing the introduction statement is a natural break point for respondents, when designing surveys, some substantive questions can be placed immediately underneath those statements on the same page (rather than presenting the introduction statement alone on a separate page). Also, survey designers can replace the long statement with a short phrase about the upcoming topic. Given that most questionnaires will have multiple topics, all these measures are believed to blur the transition from one topic to another and therefore minimise the potential impact of such transitions on breakoffs.

Secondly, based on the findings that the respondents are likely to quit the survey when they realise the extra burden, it is recommended to move the burdensome questions (e.g., filter questions) towards the end of the questionnaire. This solution is expected to delay the moment when respondents experience the burden and thus postpone the breakoff event. As a further consequence, the collected data will have fewer missing data.

In addition to optimising the survey design, a real-time intervention system can be implemented to keep respondents engaged whilst they are answering the survey. To be more specific, a gradient boosting model (or ensemble models in general) should be embedded in the back end of the survey platform as the predictive model since it achieves a good balance between the true positive and false positive rate in this study. In terms of the predictors, all types of predictors (i.e. demographics plus time-varying variables coded both concurrently and cumulatively) should be used if possible. However, given that many surveys might not have information about respondents' demographic background prior to the data collection (which is especially true for many household surveys), the concurrent coding of the time-varying variables should at least be used. This is because this specific coding can reflect the current response burden survey participants are experiencing and tends to be more predictive of the imminent breakoff event than the cumulative coding. Such a system will continuously estimate and monitor respondents' breakoff likelihood as they go through the survey questions. When their predicted breakoff likelihood exceeds the user-defined threshold, some pre-defined interventions can be triggered to keep respondents engaged (e.g., emphasising the importance of complete response, reminding respondents of the incentive conditional upon survey completion). Survey practitioners can also choose to restrict the interventions to only those types of questions that are known from prior literature to be prone to breakoffs. This restriction helps avoid the potential issue caused by the low precision commonly seen in the predictive models for breakoff (i.e. models incorrectly predict many breakoff events and trigger an excessive number of interventions throughout the entire survey). Also, when it is impossible to relocate those breakoff-prone questions to the end of the questionnaire, restricting the interventions to only those questions is a useful way to keep respondents engaged and potentially mitigate the impact of those questions on breakoff.

Even when the survey data collection ends, survey practitioners still have options to deal with the breakoff issue. The first recommended practice is to investigate whether the breakoff cases are associated with certain characteristics (e.g., socio-demographic background, and paradata recorded during the sample recruitment stage). This analysis will give an indication of the presence of the breakoff bias and its extent. If the breakoff bias is found to be present and potentially detrimental to the survey estimates, post-survey methods can be adopted to compensate for the breakoff-induced missing data. One of the recommended methods is multiple imputation. This method has been demonstrated in this study to be able to produce a low bias in the survey estimates. Also, multiple imputation retains the maximum amount of information in the data (i.e. breakoff respondents' answers to the early questions are kept in the analysis), thereby leading to higher statistical power in the analysis.

7.3 Limitations and future research

Despite the contributions of this thesis, limitations should also be pointed out. Firstly, the introduction survey breakoff needs further exploration. For the surveys used in this thesis as well as in past literature, a sizeable number of breakoffs occurred at the first few questions in the survey (referred to as introduction survey breakoff). The present thesis cannot explain the introduction survey breakoff due to the lack of frame information about these breakoff cases. Given that more paradata can be collected at the survey recruitment stage (e.g., contacts, login attempts) and this data type can indicate respondents' cooperation tendency or motivation, researchers could investigate how such data can explain the breakoff happening at the start of the survey. In addition, the frame data (and previous information collected in the longitudinal survey) can also be useful for understanding such early breakoffs.

Apart from the paradata at the survey recruitment stage, only a limited amount of paradata at the response stage is used across all three substantive chapters. Previous research has shown that the response behaviour paradata (e.g., mouse movements, change in the question response time) can help explain and predict the breakoff risk. However, the research on the coding of such paradata is still at its early stage, and most studies concentrate only on one specific paradata - question response time. Thus, research should test different coding schemes of other response behaviour paradata. Lagging should receive special attention as some researchers assume the most recent events affect the breakoff risk, but few have tested this empirically.

The study on the real-time intervention system can also be further improved. For example, all predictive models tested in this thesis have low precision. That is, out of those questions that are predicted by the model to have breakoffs, only a few of them actually have the breakoff event. Triggering interventions using a model with such a low precision level might raise many false alarms. Whether it will result in a counter effect (i.e. excessive alarms irritate respondents and cause them to break off) still needs exploration. Another limitation relating to the model is that only two survival models are fitted in this thesis because the existing software package for the survival machine learning model cannot handle the clustered data format mentioned earlier. It will be interesting to see whether the survival version of the commonly used machine learning models (e.g., survival gradient boosting, random survival forest) can outperform the classification counterparts for binary variables in the future.

In addition to the models in the real-time intervention system, the different interventions also need more research. This thesis does not touch upon this element, but a range of options are available to be tested. One option could be modularising the questionnaire to only ask those soon-to-breakoff cases the key module. This intervention ensures that at least some key information is collected, which can then be used in the post-survey adjustment. Another option is to test different content/wording of the pop-up message (e.g., emphasising the conditional incentive vs. the importance of respondents' completing the survey). This is based on the saliency-leverage theory, which states that people care about different elements in the survey request and using designs the respondents care the most about can help induce cooperation (Groves, Singer and Corning, 2000). All interventions to be investigated in the future should be based on theories regarding what motivates cooperation or reduces the response burden. Apart from this, researchers can also borrow from the existing interventions for survey nonresponse to tackle breakoff (e.g., targeting).

At the post-survey adjustment stage, the main limitation is related to the simulation. For example, all substantive answers of the breakoff cases are removed in the simulation at the same point. Yet, respondents quit the survey at different questions in practice, which leads to more diverse and complex missing data patterns among the breakoff cases. Future research can introduce an extra layer of randomness when discarding the breakoff cases' substantive answers to account for different timings of breakoff. Such a situation will likely lead to a better performance in multiple imputation as this method uses all available information during the compensation.

Another possible improvement in the simulation is to create missing data in the paradata used for the breakoff compensation. Due to technical issues or website blockers, it is common in reality to see this data type suffering from missingness. For example, the paradata about respondents' mouse back clicks and skip forwards in this thesis have so much missing data that they were discarded. Future research can also simulate the effects of missing paradata. This will provide more insight into how the breakoff compensation methods perform. Overall, all these adjustments will make the simulation more similar to real-world surveys, thereby increasing the external validity.

Additionally, when simulating breakoff, there were only two factors that uniquely affected breakoffs (i.e. responding device and average question response time). For the future studies, researchers can simulate breakoffs using more factors that only impact this survey outcome. Such setting will help investigate whether using the breakoff weighting genuinely worsens the estimates, compared to methods that do not deal with breakoff specifically,

Furthermore, no post-survey adjustment methods investigated in this thesis can solve the breakoff bias under the MNAR mechanism. This is unsurprising given that most of those methods work on the assumption of either MCAR (completely ignoring breakoff) or MAR (weighting and imputation). In the future, researchers can study those methods that are dedicated to the MNAR scenario, such as mixture models (Hedeker and Gibbons, 2006).

7.4 Final remarks

The trend of adopting web surveys will continue given their benefits in cost-saving and speed. While embracing those benefits, survey designers should devote attention to the breakoff issue in the web survey. This thesis tackles the web survey breakoff from three perspectives: (1) understanding the impact of filter question formats and question topics on breakoff and its timing, (2) predicting breakoff during the survey using the best combination of machine learning models and coding of predictors, and (3) mitigating the breakoff bias using post-survey adjustment. Applying the findings from this thesis to web survey practice will lead to survey data that have fewer missing data and more representative samples. More importantly, such data will provide governments and companies with high quality evidence for their decision-making.

References

Allen, M. (2017) ‘The SAGE encyclopedia of communication research methods’. Thousand Oaks, California: Sage Publications. Available at: https://doi.org/10.4135/9781483381411.

Bailey, J. *et al.* (2017) *Next Steps Age 25 Survey technical report*. Available at: https://cls.ucl.ac.uk/wp-content/uploads/2017/11/5545age_25_survey_questionnaire.pdf.

Bekova, S. (2021) ‘Does employment during doctoral training reduce the PhD completion rate?’, *Studies in Higher Education*, 46(6), pp. 1068–1080. Available at: https://doi.org/10.1080/03075079.2019.1672648.

Bethlehem, J. (2010) ‘Selection bias in web surveys’, *International Statistical Review*, 78(2), pp. 161–188. Available at: https://doi.org/10.1111/j.1751-5823.2010.00112.x.

Biemer, P.P. and Christ, S.L. (2008) ‘Weighting survey data’, in E.D. de Leeuw, J.J. Hox, and D.A. Dillman (eds) *International Handbook of Survey Methodology*. New York: Taylor & Francis, pp. 317–341.

Blumenberg, C. *et al.* (2018) ‘Questionnaire breakoff and item nonresponse in web-based questionnaires: Multilevel analysis of person-level and item design factors in a birth cohort’, *Journal of Medical Internet Research*, 20(12). Available at: https://doi.org/10.2196/11046.

Bodor, T. (2012) ‘Hungary’s black sunday of public opinion research: The anatomy of a failed election forecast’, *International Journal of Public Opinion Research*, 24(4), pp. 450–471. Available at: https://doi.org/10.1093/ijpor/edr039.

Bojer, C.S. and Meldgaard, J.P. (2021) ‘Kaggle forecasting competitions: An overlooked learning opportunity’, *International Journal of Forecasting*, 37(2), pp. 587–603. Available at: https://doi.org/10.1016/j.ijforecast.2020.07.007.

Bonander, C. *et al.* (2019) ‘Participation weighting based on sociodemographic register data improved external validity in a population-based cohort study’, *Journal of Clinical Epidemiology*, 108, pp. 54–63. Available at: https://doi.org/10.1016/j.jclinepi.2018.12.011.

Bosnjak, M. *et al.* (2018) ‘Establishing an open probability-based mixed-mode panel of the general population in Germany: GESIS Panel’, *Social Science Computer Review*, 36(1), pp. 103–115. Available at: https://doi.org/10.1177/0894439317697949.

Brewer, P. and Venaik, S. (2014) ‘The ecological fallacy in national culture research’, *Organization Studies*, 35(7), pp. 1063–1086. Available at: https://doi.org/10.1177/0170840613517602.

Buskirk, T.D. *et al.* (2018) ‘An introduction to machine learning methods for survey researchers’, *Survey Practice*, 11(1), pp. 1–10. Available at: https://doi.org/10.29115/SP-2018-0004.

Buskirk, T.D. (2018) ‘Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research’, *Survey Practice*, 11(1), pp. 1–13. Available at: https://doi.org/10.29115/sp-2018-0003.

Buskirk, T.D. and Kolenikov, S. (2015) ‘Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification’, *Survey Methods: Insights from the Field*, pp. 1–17. Available at: https://doi.org/10.13094/SMIF-2015-00003.

Caferra, R., Colasante, A. and Morone, A. (2021) ‘Who is afraid of the dark? Some evidence from a cross-country investigation’, *Energy Sources, Part B: Economics, Planning and Policy*, 16(11–12), pp. 1016–1025. Available at: https://doi.org/10.1080/15567249.2021.1909672.

Callegaro, M., Lozar Manfreda, K. and Vehovar, V. (2015) *Web survey methodology*. Los Angeles: SAGE.

Chen, Z. *et al.* (2022) ‘Impact of question topics and filter question formats on web survey breakoffs’, *International Journal of Market Research* [Preprint]. Available at: https://doi.org/10.1177/14707853211068008.

Christoph, M. (2019) *Interpretable machine learning: A guide for making black box models explainable*. lulu.

Clark-Fobia, A., Kephart, K. and Nelson, D. V. (2018) ‘A qualitative study on the effects of grouped versus interleafed filter questions’, *Survey Practice*, 11(2), pp. 1–11. Available at: https://doi.org/10.29115/sp-2018-0009.

Conrad, F. *et al.* (2017) ‘Reducing speeding in web surveys by providing immediate feedback’, *Survey Research Methods*, 11(1), pp. 45–61. Available at: https://doi.org/10.18148/srm/2017.v11i1.6304.

Conrad, F.G. *et al.* (2010) ‘The impact of progress indicators on task completion’, *Interacting with Computers*, 22(5), pp. 417–427. Available at: https://doi.org/10.1016/j.intcom.2010.03.001.

Cornesse, C. *et al.* (2020) ‘A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research’, *Journal of Survey Statistics and Methodology*, 8(1), pp. 4–36. Available at: https://doi.org/10.1093/jssam/smz041.

Couper, M.P. (2000) ‘Web surveys: A review of issues and approaches’, *Public Opinion Quarterly*, 64(4), pp. 464–494. Available at: https://doi.org/10.1086/318641.

CRONOS team (2018) *CROss-National Online Survey (CRONOS) panel: Data and documentation user guide*. Available at: http://www.europeansocialsurvey.org/docs/cronos/CRONOS_user_guide_e01_1.pdf.

Curley, C. *et al.* (2019) ‘Dealing with missing data: A comparative exploration of approaches using the Integrated City Sustainability Database’, *Urban Affairs Review*, 55(2), pp. 591–615. Available at: https://doi.org/10.1177/1078087417726394.

Daikeler, J., Bosnjak, M. and Manfreda, K.L. (2020) ‘Web versus other survey modes: An updated and extended meta-analysis comparing response rates’, *Journal of Survey Statistics and Methodology*, 8(3), pp. 513–539. Available at: https://doi.org/10.1093/jssam/smz008.

Department for Digital, Culture, Media & Sport. (2022) *DCMS Participation Survey 2021/22: Technical Note Quarter 1 (October-December 2021)*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1094813/DCMS_Participation_Survey_2021-22_-_Q1_Technical_Note_V3.pdf.

Durrant, G.B. and Steele, F. (2009) ‘Multilevel modelling of refusal and non-contact in household surveys: Evidence from six UK Government surveys’, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(2), pp. 361–381. Available at: https://doi.org/10.1111/j.1467-985X.2008.00565.x.

Eckman, S. *et al.* (2014) ‘Assessing the mechanisms of misreporting to filter questions in surveys’, *Public Opinion Quarterly*, 78(3), pp. 721–733. Available at: https://doi.org/10.1093/poq/nfu030.

Eckman, S. (2021) ‘Underreporting of purchases in the US Consumer Expenditure Survey’, *Journal of Survey Statistics and Methodology*, pp. 1–24. Available at: https://doi.org/10.1093/jssam/smab024.

Eckman, S. and Kreuter, F. (2018) ‘Misreporting to looping questions in surveys: Recall, motivation and burden’, *Survey Research Methods*, 12(1), pp. 59–74. Available at: https://doi.org/10.18148/srm/2018.v12i1.7168.

Enders, C.K. (2010) *Applied missing data analysis*. 2nd edn. New York: Guilford Press.

Fernández-Fontelo, A. *et al.* (2021) 'Predicting question difficulty in web surveys: A machine learning approach based on mouse movement features', *Social Science Computer Review*, pp. 1–22. Available at: https://doi.org/10.1177/08944393211032950.

Funke, F., Reips, U.-D. and Thomas, R.K. (2011) 'Sliders for the smart: Type of rating scale on the web interacts with educational level', *Social Science Computer Review*, 29(2), pp. 221–231. Available at: https://doi.org/10.1177/0894439310376896.

Galesic, M. (2006) 'Dropouts on the web: Effects of interest and burden experienced during an online survey', *Journal of Official Statistics*, 22(2), pp. 313–328.

Groves, R.M. (2011) 'Three eras of survey research', *Public Opinion Quarterly*, 75(5), pp. 861–871. Available at: https://doi.org/10.1093/poq/nfr057.

Groves, R.M., Presser, S. and Dipko, S. (2004) 'The role of topic interest in survey participation decisions', *Public Opinion Quarterly*, 68(1), pp. 2–31. Available at: https://doi.org/10.1093/poq/nfh002.

Groves, R.M., Singer, E. and Corning, A. (2000) 'Leverage-saliency theory of survey participation: description and an illustration', *Public Opinion Quarterly*, 64(3), pp. 299–308. Available at: https://doi.org/10.1086/317990.

Harrell, F.E., Lee, K.L. and Mark, D.B. (1996) 'Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicine*, 15(4), pp. 361–387. Available at: https://doi.org/https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction*. 2nd edn. New York, NY: Springer. Available at: https://doi.org/10.1007/978-0-387-84858-7.

Healey, B. (2007) 'Drop downs and scroll mice: The effect of response option format and input mechanism employed on data quality in web surveys', *Social Science Computer Review*, 25(1), pp. 111–128. Available at: https://doi.org/10.1177/0894439306293888.

Hedeker, D. and Gibbons, R.D. (2006) 'Missing data in longitudinal studies', in *Longitudinal Data Analysis*. Hoboken, New Jersey: John Wiley & Sons, pp. 279–312. Available at: https://doi.org/10.1002/0470036486.ch14.

Hochheimer, C.J. *et al.* (2016) 'Methods for evaluating respondent attrition in web-based surveys', *Journal of Medical Internet Research*, 18(11), pp. 1–11. Available at: https://doi.org/10.2196/jmir.6342.

Hoerger, M. (2010) ‘Participant dropout as a function of survey length in internet-mediated university studies: Implications for study design and voluntary participation in psychological research’, *Cyberpsychology, Behavior and Social Networking*, 13(6), pp. 697–700. Available at: https://doi.org/10.1089/cyber.2009.0445.

Hohne, J.K., Schlosser, S. and Krebs, D. (2017) ‘Investigating cognitive effort and response quality of question formats in web surveys using paradata’, *Field Methods*, 29(4), pp. 365–382. Available at: https://doi.org/10.1177/1525822X17710640.

Horwitz, R., Kreuter, F. and Conrad, F. (2017) ‘Using mouse movements to predict web survey response difficulty’, *Social Science Computer Review*, 35(3), pp. 388–405. Available at: https://doi.org/10.1177/0894439315626360.

James, G. *et al.* (2013) *An introduction to statistical learning: with applications in R*. New York, NY: Springer. Available at: https://doi.org/10.1007/978-1-4614-7138-7.

Kern, C., Klausch, T. and Kreuter, F. (2019) ‘Tree-based machine learning methods for survey research’, *Survey Research Methods*, 13(1), pp. 73–93. Available at: https://doi.org/10.18148/srm/2019.v13i1.7395.

Kern, C., Weiß, B. and Kolb, J.-P. (2021) ‘Predicting nonresponse in future waves of a probability-based mixed-mode panel with machine learning’, *Journal of Survey Statistics and Methodology*, pp. 1–24. Available at: https://doi.org/10.1093/jssam/smab009.

Kirchner, A. and Signorino, C.S. (2018) ‘Using support vector machines for survey research’, *Survey Practice*, 11(1), pp. 1–14. Available at: https://doi.org/10.29115/SP-2018-0001.

Klein, D.J. *et al.* (2011) ‘Understanding nonresponse to the 2007 Medicare CAHPS survey’, *Gerontologist*, 51(6), pp. 843–855. Available at: https://doi.org/10.1093/geront/gnr046.

Kreuter, F. *et al.* (2011) ‘The effects of asking filter questions in interleafed versus grouped format’, *Sociological Methods and Research*, 40(1), pp. 88–104. Available at: https://doi.org/10.1177/0049124110392342.

Kreuter, F. (2013) ‘Improving surveys with paradata: Introduction’, in F. Kreuter (ed.) *Improving surveys with paradata: Analytic uses of process information*. Hoboken, New Jersey: John Wiley & Sons, pp. 1–9. Available at: https://doi.org/10.1002/9781118596869.ch1.

Kreuter, F. (2017) ‘Getting the most out of paradata’, in D.L. Vannette, J.A. Krosnick, and J.W. Sakshaug (eds) *The Palgrave handbook of survey research*. Basingstoke: Palgrave Macmillan, pp. 193–198. Available at: https://doi.org/doi.org/10.1007/978-3-319-54395-6_24.

Kreuter, F., Eckman, S. and Tourangeau, R. (2020) 'The salience of survey burden and its effect on response behavior to skip questions: Experimental results from telephone and web surveys', in P. Beatty et al. (eds) *Advances in questionnaire design, development, evaluation and testing*. Hoboken, NY: John Wiley & Sons, pp. 213–227. Available at: https://doi.org/10.1002/9781119263685.ch9.

Kreuter, F., Presser, S. and Tourangeau, R. (2008) 'Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity', *Public Opinion Quarterly*, 72(5), pp. 847–865. Available at: https://doi.org/10.1093/poq/nfn063.

Kuhn, M. and Johnson, K. (2013) *Applied predictive modeling*. New York: Springer. Available at: https://doi.org/10.1007/978-1-4614-6849-3.

Lantz, B. (2019) *Machine learning with R: Expert techniques for predictive modeling*. 3rd edn. Birmingham, UK: Packt.

Lavrakas, P.J. (2008) *Encyclopedia of survey research methods*. Thousand Oaks, California: Sage Publications. Available at: https://doi.org/10.4135/9781412963947.

Lee, S. and Lim, H. (2019) 'Review of statistical methods for survival analysis using genomic data', *Genomics and Informatics*, 17(4). Available at: https://doi.org/10.5808/GI.2019.17.4.e41.

Leon, A.C. *et al.* (2006) 'Attrition in randomized controlled clinical trials: Methodological issues in psychopharmacology', *Biological Psychiatry*, 59(11), pp. 1001–1005. Available at: https://doi.org/10.1016/j.biopsych.2005.10.020.

Liu, M. (2020) 'Using machine learning models to predict attrition in a survey panel', in C.A. Hill et al. (eds) *Big Data meets survey science: A collection of innovative methods*. Hoboken, NY: John Wiley & Sons, pp. 415–433. Available at: https://doi.org/10.1002/9781118976357.ch14.

Liu, M. and Wronski, L. (2018) 'Examining completion rates in web surveys via over 25,000 real-world surveys', *Social Science Computer Review*, 36(1), pp. 116–124. Available at: https://doi.org/10.1177/0894439317695581.

Lugtig, P. and Luiten, A. (2021) 'Do shorter stated survey length and inclusion of a QR code in an invitation letter lead to better response rates?', *Survey Methods: Insights from the Field* [Preprint]. Available at: https://doi.org/10.13094/SMIF-2021-00001.

Matzat, U., Snijders, C. and van der Horst, W. (2009) 'Effects of different types of progress indicators on drop-out rates in web surveys', *Social Psychology*, 40(1), pp. 43–52. Available at: https://doi.org/10.1027/1864-9335.40.1.43.

Mavletova, A. and Couper, Mike.P. (2015) ‘A meta-analysis of breakoff rates in mobile web surveys’, in D. Toninelli, R. Pinter, and P. de Pedraza (eds) *Mobile research methods: Opportunities and challenges of mobile research methodologies*. London: Ubiquity Press, pp. 81–98. Available at: https://doi.org/10.5334/bar.f.

Mayr, A. *et al.* (2014) ‘The evolution of boosting algorithms: from machine learning to statistical modelling’, *Methods of Information in Medicine*, 53(6), pp. 419–427. Available at: https://doi.org/10.3414/ME13-01-0122.

McClain, C.A. *et al.* (2018) ‘A typology of web survey paradata for assessing total survey error’, *Social Science Computer Review*, pp. 1–18. Available at: https://doi.org/10.1177/0894439318759670.

McCoy, T.P. *et al.* (2009) ‘Attrition bias in a U.S. internet survey of alcohol use among college freshmen’, *Journal of Studies on Alcohol and Drugs*, 70(4), pp. 606–614. Available at: https://doi.org/10.15288/jsad.2009.70.606.

McGonagle, K.A. (2013) ‘Survey breakoffs in a computer-assisted telephone interview’, *Survey Research Methods*, 7(2), pp. 79–90. Available at: https://doi.org/10.18148/srm/2013.v7i2.5126.

Mills, M. (2011) *Introducing survival and event history analysis*. London: SAGE.

Mittereder, F. (2022) ‘Breakoff in Web Surveys’, in P. Atkinson et al. (eds) *Sage research methods foundations*. London: SAGE Publications. Available at: https://doi.org/https://dx.doi.org/10.4135/9781526421036927734.

Mittereder, F. and West, B.T. (2021) ‘A dynamic survival modeling approach to the prediction of web survey breakoff’, *Journal of Survey Statistics and Methodology*, pp. 1–34. Available at: https://doi.org/10.1093/jssam/smab015.

Mittereder, F.K. (2019) *Predicting and preventing breakoff in web surveys*. University of Michigan.

Morales, D.X., Morales, S.A. and Beltran, T.F. (2021) ‘Racial/Ethnic disparities in household food insecurity during the COVID-19 pandemic: A nationally representative study’, *Journal of Racial and Ethnic Health Disparities*, 8(5), pp. 1300–1314. Available at: https://doi.org/10.1007/s40615-020-00892-7.

Muggah, E.M. and McSweeney, M.B. (2017) ‘Females’ attitude and preference for beer: A conjoint analysis study’, *International Journal of Food Science and Technology*, 52(3), pp. 808–816. Available at: https://doi.org/10.1111/ijfs.13340.

Natekin, A. and Knoll, A. (2013) ‘Gradient boosting machines, a tutorial’, *Frontiers in Neurorobotics*, 7. Available at: https://doi.org/10.3389/fnbot.2013.00021.

Nissen, J., Donatello, R. and Van Dusen, B. (2019) ‘Missing data and bias in physics education research: A case for using multiple imputation’, *Physical Review Physics Education Research*, 15(2), pp. 1–15. Available at: https://doi.org/10.1103/PhysRevPhysEducRes.15.020106.

Noel, H.J. and Huang, A.R. (2019) ‘The effect of varying incentive amounts on physician survey response’, *Evaluation and the Health Professions*, 42(1), pp. 71–81. Available at: https://doi.org/10.1177/0163278718809844.

Nowok, B., Raab, G.M. and Dibben, C. (2016) ‘Synthpop: Bespoke creation of synthetic data in R’, *Journal of Statistical Software*, 74(11). Available at: https://doi.org/10.18637/jss.v074.i11.

Olson, K., Smyth, J.D. and Wood, H.M. (2012) ‘Does giving people their preferred survey mode actually increase survey participation rates? An experimental examination’, *Public Opinion Quarterly*, 76(4), pp. 611–635. Available at: https://doi.org/10.1093/poq/nfs024.

Peytchev, A. (2009) ‘Survey breakoff’, *Public Opinion Quarterly*, 73(1), pp. 74–97. Available at: https://doi.org/10.1093/poq/nfp014.

Peytchev, A. (2011) ‘Breakoff and unit nonresponse across web surveys’, *Journal of Official Statistics*, 27(1), pp. 33–47.

R Core Team (2020) ‘R: A language and environment for statistical computing’. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team (2021) ‘R: A language and environment for statistical computing’. Vienna, Austria: R Foundation for Statistical Computing.

Revilla, M. (2017) ‘Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain’, *methods, data, analyses*, 11(2), pp. 135–162. Available at: https://doi.org/10.12758/mda.2017.02.

Rhys, H.I. (2020) *Machine Learning with R, the tidyverse, and mlr*. Shelter Island, New York: Manning Publications.

Roßmann, J., Gummer, T. and Silber, H. (2018) ‘Mitigating satisficing in cognitively demanding grid questions: evidence from two web-based experiments’, *Journal of Survey Statistics and Methodology*, 6(3), pp. 376–400. Available at: https://doi.org/10.1093/jssam/smx020.

Roster, C.A., Albaum, G. and Smith, S.M. (2017) ‘Effect of topic sensitivity on online survey panelists’ motivation and data quality’, *Journal of Marketing Theory and Practice*, 25(1), pp. 1–16. Available at: https://doi.org/10.1080/10696679.2016.1205449.

Rubin, D.B. (1976) ‘Inference and missing data’, *Biometrika*, 63(3), pp. 581–592. Available at: https://doi.org/10.2307/2335739.

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. New York: John Wiley and Sons.

Sakshaug, J.W. and Crawford, S.D. (2010) ‘The impact of textual messages of encouragement on web survey breakoffs: An experiment’, *International Journal of Internet Science*, 4(1), pp. 50–60.

Scherpenzeel, A. (2011) ‘Data collection in a probability-based internet panel: How the LISS panel was built and how it can be used’, *Bulletin de Méthodologie Sociologique*, 109(1), pp. 56–61. Available at: https://doi.org/10.1177/0759106310387713.

Schober, P. and Vetter, T.R. (2018) ‘Survival analysis and interpretation of time-to-event data: The tortoise and the hare’, *Anesthesia & Analgesia*, 127(3), pp. 792–798. Available at: https://doi.org/10.1213/ANE.0000000000003653.

Shropshire, K.O., Hawdon, J.E. and Witte, J.C. (2009) ‘Web survey design: Balancing measurement, response, and topical interest’, *Sociological Methods and Research*, 37(3), pp. 344–370. Available at: https://doi.org/10.1177/0049124108327130.

Signorino, C.S. and Kirchner, A. (2018) ‘Using LASSO to model interactions and nonlinearities in survey data’, *Survey Practice*, 11(1), pp. 1–10. Available at: https://doi.org/10.29115/SP-2018-0005.

Silber, H., Lischewski, J. and Leibold, J. (2013) ‘Comparing different types of web surveys: examining drop-outs, non-response and social desirability’, *Metodoloski Zvezki*, 10(2), pp. 121–143.

Singer, J.D. and Willett, J.B. (2003) *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.

Sischka, P.E. *et al.* (2022) ‘The impact of forced answering and reactance on answering behavior in online surveys’, *Social Science Computer Review*, 40(2), pp. 405–425. Available at: https://doi.org/10.1177/0894439320907067.

Spooner, A. *et al.* (2020) ‘A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction’, *Scientific Reports*, 10. Available at: https://doi.org/10.1038/s41598-020-77220-w.

Steinbrecher, M., Roßmann, J. and Blumenstiel, J.E. (2015) ‘Why do respondents break off web surveys and does it matter? Results from four follow-up surveys’, *International Journal of Public Opinion Research*, 27(2), pp. 289–302. Available at: https://doi.org/10.1093/ijpor/edu025.

Stern, M.J. (2008) ‘The use of client-side paradata in analyzing the effects of visual layout on changing responses in web surveys’, *Field Methods*, 20(4), pp. 377–398. Available at: https://doi.org/10.1177/1525822X08320421.

Sterne, J.A.C. *et al.* (2009) ‘Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls’, *British Medical Journal*, 338(7713), pp. 157–160. Available at: https://doi.org/10.1136/bmj.b2393.

Stoltzfus, J.C. (2011) ‘Logistic regression: A brief primer’, *Academic Emergency Medicine*, 18(10), pp. 1099–1104. Available at: https://doi.org/10.1111/j.1553-2712.2011.01185.x.

Teclaw, R., Price, M.C. and Osatuke, K. (2012) ‘Demographic question placement: Effect on item response rates and means of a veterans health administration survey’, *Journal of Business and Psychology*, 27(3), pp. 281–290. Available at: https://doi.org/10.1007/s10869-011-9249-y.

Therneau, M.T. (2021) ‘A package for survival analysis in R’. Available at: https://cran.r-project.org/package=survival.

Tibshirani, R. (1997) ‘The LASSO method for variable selection in the Cox model’, *Statistics in Medicine*, 16(4), pp. 385–395. Available at: https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

Tijdens, K. (2014) ‘Dropout rates and response times of an occupation search tree in a web survey’, *Journal of Official Statistics*, 30(1), pp. 23–43. Available at: https://doi.org/10.2478/jos-2014-0002.

Toepoel, V. (2015) *Doing surveys online*. London: SAGE Publications.

Toepoel, V. and Lugtig, P. (2018) ‘Modularization in an era of mobile web: Investigating the effects of cutting a survey into smaller pieces on data quality’, *Social Science Computer Review* [Preprint]. Available at: https://doi.org/10.1177/0894439318784882.

Torcal, M. and Christmann, P. (2021) ‘Responsiveness, performance and corruption: Reasons for the decline of political trust’, *Frontiers in Political Science*, 3(July), pp. 1–13. Available at: https://doi.org/10.3389/fpos.2021.676672.

Tourangeau, R. (2018) ‘The survey response process from a cognitive viewpoint’, *Quality Assurance in Education*, 26(2), pp. 169–181. Available at: https://doi.org/10.1108/qae-06-2017-0034.

Tourangeau, R., Conrad, F. and Couper, M.P. (2013) *The science of web surveys*. New York: Oxford University Press.

Tourangeau, R. and Yan, T. (2007) ‘Sensitive questions in surveys’, *Psychological Bulletin*, 133(5), pp. 859–883. Available at: https://doi.org/10.1037/0033-2909.133.5.859.

United States Census Bureau (2022) *Quick facts*. Available at: https://www.census.gov/quickfacts/fact/table/US/PST045222 (Accessed: 17 January 2023).

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) 'mice: Multivariate imputation by chained equations in R', *Journal of Statistical Software*, 45(3), pp. 1–67. Available at: https://doi.org/10.18637/jss.v045.i03.

van der Meer, T. (2010) 'In what we trust? A multi-level study into trust in parliament as an evaluation of state characteristics', *International Review of Administrative Sciences*, 76(3), pp. 517–536. Available at: https://doi.org/10.1177/0020852310372450.

Vehovar, V. and Cehovin, G. (2014) 'Questionnaire length and breakoffs in web surveys: A meta study', in *Seventh Internet Survey Methodology Workshop 2014*. Available at: http://www.websm.org/db/12/17719/.

Villar, A., Callegaro, M. and Yang, Y. (2013) 'Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys', *Social Science Computer Review*, 31(6), pp. 744–762. Available at: https://doi.org/10.1177/0894439313497468.

Villar, A. and Sommer, E. (2017) *Web recruitment design plans and experimental testing*. Available at: https://seriss.eu/wp-content/uploads/2017/12/SERISS-Deliverable-7.3-Web-recruitment-design-plans-and-experimental-testing.pdf.

Vink, G. *et al.* (2014) 'Predictive mean matching imputation of semicontinuous variables', *Statistica Neerlandica*, 68(1), pp. 61–90. Available at: https://doi.org/10.1111/stan.12023.

Wang, P., Li, Y. and Reddy, C.K. (2019) 'Machine learning for survival analysis: A survey', *ACM Computing Surveys*, 51(6). Available at: https://doi.org/10.1145/3214306.

Wenz, A. (2017) *Sources of error in mobile survey data collection*. Institute for Social and Economic Research, University of Essex.

Willett, J.B. and Singer, J.D. (1993) 'It's about time: Using discrete-time survival analysis to study duration and the timing of events', *Journal of Educational Statistics*, 18(2), pp. 155–195.

Zhang, C. and Conrad, F.G. (2014) 'Speeding in web surveys: The tendency to answer very fast and its association with straightlining', *Survey Research Methods*, 8(2), pp. 127–135. Available at: https://doi.org/10.18148/srm/2014.v8i2.5453.

Appendices

Appendix A

Table A.1 Descriptive summary of categorical variables used in the analysis, before and after multiple imputation.

	Before imputation		After imputation		
Variable	**Frequency**	**Percent**	**Frequency**	**Percent**	**Source**[a]
Respondent-level[b] *(N = 3128)*					
Breakoff					Survey
No	2608	83.38	2608	83.38	
Yes	520	16.62	520	16.62	
Education					Survey
High school or below	700	22.38	901	28.81	
College	964	30.82	983	31.42	
Bachelor and above	1219	38.97	1244	39.77	
Missing	245	7.83	-	-	
Ethnicity					Recruitment
Non-white	705	22.54	740	23.67	
White	2301	73.56	2388	76.33	
Missing	122	3.90	-	-	
Gender					Recruitment
Female	2060	65.86	2093	66.92	
Male	1019	32.58	1035	33.08	
Missing	49	1.57	-	-	
Household income					Recruitment
Low	774	24.74	787	25.16	
Middle	1468	46.93	1493	47.73	
High	837	26.76	848	27.11	
Missing	49	1.57	-	-	
Current student					Survey
No	2169	69.34	2192	70.08	
Yes	704	22.51	936	29.92	
Missing	255	8.15	-	-	
Marital status					Survey

No	1458	46.61	1637	52.32	
Yes	1408	45.01	1491	47.68	
Missing	262	8.38	-	-	
Responding device					Survey
Non-mobile device	1896	60.61	1896	60.61	
Mobile device	1229	39.29	1229	39.29	
Missing	3	0.10	3	0.10	
Number of sessions					Survey
One session	2961	94.66	2961	94.66	
More than one session	167	5.34	167	5.34	
Filter question format					Survey
Grouped	1544	49.36	1544	49.36	
Interleafed	1584	50.64	1584	50.64	
Question-level[c] *(N = 196)*					
Topic					Paradata
Introduction statement	6	3.06	6	3.06	
Demographic	9	4.59	9	4.59	
Housing	10	5.10	10	5.10	
Clothing	72	36.73	72	36.73	
Utilities	30	15.31	30	15.31	
Insurance	54	27.55	54	27.55	
Income	15	7.65	15	7.65	
Matrix question					Paradata
No	194	98.98	194	98.98	
Yes	2	1.02	2	1.02	
Open-ended question					Paradata
No	125	63.78	125	63.78	
Yes	71	36.22	71	36.22	

[a] Variables come from three sources: (1) from the survey itself; (2) recorded during the recruitment stage; (3) paradata

[b] All respondent-level variables are time-constant

[c] All question-level variables are time-varying

Table A.2 Descriptive summary of continuous variables used in the analysis.

Variable	Min	Max	Median	Mean	SD	N	Missing (%)	Source[a]
Respondent-level[b]								
Age[c]	18	81.00	35.00	43.80	18.20	3128	1.57	Recruitment
Survey duration (min)	0	123.88	9.82	12.18	11.09	3128	0	Survey
Question-level[d]								
Number of questions seen	1	187.00	79.00	73.51	33.09	3128	0	Paradata
Item nonresponse rate	0	100.00	7.69	14.01	18.85	3128	0	Paradata
Word count of question stems	1	55.00	15.50	15.66	10.19	196	0	Paradata

[a] Variables come from three sources: (1) from the survey itself; (2) recorded during the recruitment stage; (3) paradata

[b] All respondent-level variables are time-constant

[c] The descriptive summary for this variable is the same before and after the imputation

[d] All question-level variables are time-varying

Table A.3 Odds ratio of logistic regression predicting breakoffs where missing data is categorised separately - based on the quadratic time specification and the full sample size.

Variable	Model 1	Model 2
Intercept	0.004***	0.13***
Number of questions seen (linear)	0.95***	0.93***
Number of questions seen (quadratic)	1.0002**	1.0004***
Married (ref: no)		
Yes	0.76**	0.87
Missing	3.65***	2.06***
Male (ref: female)	1.06	1.03
Age	1.01**	1.01***
Non-white (ref: white)		
Yes	0.74**	0.92
Missing	0.44***	0.54**
Current student (ref: no)		
Yes	1.16	0.95
Missing	6.19***	3.13***
Education (ref: high school or below)		
College	1.00	0.98
Bachelor or above	0.78	0.77*
Missing	7.22***	4.49***
Household income (ref: low)		
Middle	1.06	1.10
High	1.24	1.35*
Topic (ref: Introduction Statement)		
Demographics		0.05***
Housing		0.34***
Clothing		0.35***
Utilities		0.36***
Insurance		0.58**
Income		0.06***
Matrix question (ref: no)		1.37
Open-ended question (ref: no)		0.85
Question stem word count		0.98***
Item nonresponse rate		1.002
Grouped (ref: Interleafed)		1.09
Mobile device (ref: non-mobile)		1.15
Multiple sessions (ref: one session)		0.72
Survey duration (min)		0.80***
N of Respondents	3,078	3,078
N of Observations	226,213	226,213
Log Likelihood	-2,516.79	-2,234.43
AIC	5,065.59	4,528.85

* p < 0.1, ** p < 0.05, *** p < 0.01

Table A.4 Odds ratio of logistic regression predicting breakoffs where missing data is categorised separately - based on the quadratic time specification and the restricted sample.

Variable	**Model 3**	**Model 4**
Intercept	0.10***	0.18*
Number of questions seen (linear)	0.91***	0.88***
Number of questions seen (quadratic)	1.0006***	1.0007**
Married (ref: no)		
Yes	1.08	1.08
Missing	3.78***	3.64***
Male (ref: female)	0.92	0.92
Age	0.998	0.998
Non-white (ref: white)		
Yes	0.96	0.97
Missing	1.25	1.30
Current student (ref: no)		
Yes	0.95	0.95
Missing	0.84	0.88
Education (ref: high school or below)		
College	0.97	0.97
Bachelor or above	0.75	0.76
Missing	0.000001	0.000001
Household income (ref: low)		
Middle	0.97	0.97
High	1.68**	1.68**
Topic (ref: Clothing)		
Utilities	1.10	0.86
Insurance	1.90***	3.98
Introduction Statement	3.00***	3.32
Open-ended question (ref: no)	0.45***	0.44***
Question stem word count	0.98*	0.99
Item nonresponse rate	0.9998	1.0001
Grouped (ref: Interleafed)	1.17	0.18***
Mobile device (ref: non-mobile)	1.32	1.32*
Multiple sessions (ref: one session)	1.00	1.01
Survey duration (min)	0.85***	0.85***
Grouped x Questions seen (linear)		1.08***
Grouped x Questions seen (quadratic)		0.9994**
Utilities x Questions seen (linear)		1.01
Utilities x Questions seen (quadratic)		0.99995
Insurance x Questions seen (linear)		0.97
Insurance x Questions seen (quadratic)		1.0003
Introduction Statement x Questions seen (linear)		0.99
Introduction Statement x Questions seen (quadratic)		1.0001
N of Respondents	2,754	2,754
N of Observations	146,731	146,731
Log Likelihood	-1,227.25	-1,220.43
AIC	2,506.51	2,508.86

* p < 0.1, ** p < 0.05, *** p < 0.01

Table A.5 Log odds (standard errors) of logistic regression predicting breakoffs using different time specifications.

	Linear	**Quadratic**	**Cubic**	**4th**	**5th**
Time	**Time**	**Time**	**Time**	**Power**	**Power**
Number of questions seen	-0.54***	-0.61***	-0.99***	-1.15***	-1.68***
	(0.006)	(0.007)	(0.01)	(0.02)	(0.03)
Number of questions seen^2		0.003***	0.03***	0.04***	0.11***
		(0.00004)	(0.0006)	(0.0009)	(0.002)
Number of questions seen^3			-0.0002***	-0.0006***	-0.003***
			(0.000006)	(0.00001)	(0.00008)
Number of questions seen^4				0.000002***	0.00003***
				(0.00000006)	(0.000001)
Number of questions seen^5					-0.0000001***
					(0.000000004)
N of Respondents	3,128	3,128	3,128	3,128	3,128
N of Observations	229,940	229,940	229,940	229,940	229,940
Log Likelihood	-8,024.00	-7,285.63	-5,122.42	-4,729.78	-3,915.69
AIC	16,049.99	14,575.25	10,250.84	9,467.55	7,841.39

* p<0.1; ** p<0.05; *** p<0.01

Table A.6 Goodness-of-fit of logistic regression predicting breakoffs using different time specifications.

Time specification	**Deviance**	**AIC**	**BIC**	**Deviance difference**	**AIC difference**	**BIC difference**
Linear	16,048	16,050	16,060	-	-	-
Quadratic	14,571	14,575	14,596	1,477	1,475	1,464
Cubic	10,245	10,251	10,282	4,326	4,324	4,314
4th Power	9,460	9,468	9,509	785	783	773
5th Power	7,831	7,841	7,893	1,628	1,626	1,616

Table A.7 Odds ratio of logistic regression predicting breakoffs where the reference category for the question topic is Insurance - based on the quadratic time specification and the restricted sample.

Variable	**Model 3**	**Model 4**
Intercept	0.21***	0.82
Number of questions seen (linear)	0.91***	0.85***
Number of questions seen (quadratic)	1.0006***	1.001***
Married (ref: no)	1.10	1.10
Male (ref: female)	0.97	0.97
Age	0.998	0.998
Non-white (ref: white)	0.96	0.97
Current Student (ref: no)	0.94	0.94
Education (ref: high school or below)		
College	0.96	0.96
Bachelor or above	0.76	0.77
Household income (ref: low)		
Middle	0.96	0.96
High	1.63**	1.63**
Topic (ref: Insurance)		
Clothing	0.58***	0.23*
Utilities	0.62**	0.20*
Introduction Statement	1.72**	0.99
Open-ended question (ref: no)	0.42***	0.42***
Question stem word count	0.98**	0.98**
Item nonresponse rate	1.001	1.001
Grouped (ref: Interleafed)	1.19	0.19***
Mobile device (ref: non-mobile)	1.38**	1.39**
Multiple sessions (ref: one session)	0.95	0.95
Survey duration (min)	0.84***	0.85***

Grouped x Questions seen (linear)		1.08***
Grouped x Questions seen (quadratic)		0.9995**
Clothing x Questions seen (linear)		1.04
Clothing x Questions seen (quadratic)		0.9997
Utilities x Questions seen (linear)		1.05
Utilities x Questions seen (quadratic)		0.9996
Introduction Statement x Questions seen (linear)		1.02
Introduction Statement x Questions seen (quadratic)		0.9998
N of Respondents	2,798	2,798
N of Observations	149,154	149,154
Log Likelihood	-1,269.67	-1,262.49
AIC	2,583.35	2,584.98

* p < 0.1, ** p < 0.05, *** p < 0.01

Appendix B

Table B.1 Descriptive summary of categorical variables used in the analysis, by survey wave.

	Wave 1		Wave 2	
Variable	**Frequency**	**Percent**	**Frequency**	**Percent**
Respondent-level (N = 3128)			*Respondent-level (N = 2370)*	
Breakoff				
No	2608	83	1967	83
Yes	520	17	403	17
Education				
High school or below	700	22	501	21
College	964	31	725	31
Bachelor and above	1219	39	879	37
Missing	245	8	265	11
Ethnicity				
Non-white	705	23	387	16
White	2301	74	1746	74
Missing	122	4	237	10
Current student				
No	2169	69	1703	72
Yes	704	23	402	17
Missing	255	8	265	11
Marital status				
No	1458	47	1008	43
Yes	1408	45	1087	46
Missing	262	8	275	12
Responding device				
Non-mobile device	1896	61	1663	70
Mobile device	1229	39	693	29
Missing	3	0	14	1
Number of sessions				
One session	2961	95	2308	97
More than one session	167	5	62	3
Filter question format				

Grouped	1544	49	1212	51
Interleafed	1584	51	1158	49
Question order (within randomised blocks)				
High-low frequency	-	-	1161	49
Low-high frequency	-	-	1209	51
Question-level (N = 196)			*Question-level (N = 126)*	
Topic				
Introduction	6	3	6	5
Demographic	9	5	8	6
Housing	10	5	4	3
Clothing	72	37	36	29
Utilities	30	15	30	24
Insurance	54	28	30	24
Income	15	8	12	10
Introduction statement				
No	190	96.94	120	95.24
Yes	6	3.06	6	4.76
Matrix question				
No	194	99	124	98
Yes	2	1	2	2
Open-ended question				
No	125	64	91	72
Yes	71	36	35	28

Table B.2 Descriptive summary of continuous variables used in the analysis, by survey wave.

Variable	Wave	Min	Max	Median	Mean	SD	N	Missing (%)
Respondent-level								
Age	1	18	81	35	43.80	18.20	3128	1.57
	2	18	81	49	48.29	17.18	2370	10.42
Question-level								
Number of questions seen [a]	1	1	187	79	73.50	33.11	3128	0
	2	1	119	64	58.11	26.57	2370	0
Item nonresponse rate (%) [a]	1	0	100	7.89	14.03	18.83	3128	0
	2	0	100	5.88	12.41	22.99	2370	0
Word count of question stems	1	1	55	15.50	15.66	10.19	196	0
	2	1	55	14.00	15.85	10.82	126	0

[a] This variable is recorded at the question level but aggregated here to the respondent level for the purpose of description only.

Table B.3 Best hyperparameter values of classification models.

Model type	Predictor group	Hyperparameter
Logistic	Demographics	-
	Concurrent	-
	Cumulative	-
	All combined	-
LASSO logistic	Demographics	λ = 0.00901
	Concurrent	λ = 0.00077
	Cumulative	λ = 0.00301
	All combined	λ = 0.00670
SVM	Demographics	C = 7.37 $\varphi(\cdot)$ = radial Sigma = 0.0020
	Concurrent	C = 9.78 $\varphi(\cdot)$ = radial Sigma = 0.0057
	Cumulative	C = 17.9 $\varphi(\cdot)$ = radial Sigma = 0.0091
	All combined	C = 2.18 $\varphi(\cdot)$ = radial Sigma = 0.0102

Random forest	Demographics	mtry = 5
		trees = 1970
		min_n = 36

	Concurrent	mtry = 4
		trees = 1676
		min_n = 17

	Cumulative	mtry = 2

		trees = 1072
		min_n = 38

	All combined	mtry = 5
		trees = 1402
		min_n = 15

Boosting	Demographics	mtry = 10
		trees = 651
		min_n = 5
		tree depth = 6
		learn_rate = 0.00178

	Concurrent	mtry = 3
		trees = 909
		min_n = 15
		tree depth = 15
		learn_rate = 0.0549

	Cumulative	mtry = 5
		trees = 471
		min_n = 8
		tree depth = 5
		learn_rate = 0.00677

	All combined	mtry = 15
		trees = 471
		min_n = 8
		tree depth = 5
		learn_rate = 0.00677

Appendix C

The boxplots below show the distribution of the survey estimates. The plots are separated by breakoff rates (right, row) and breakoff mechanisms (column). Each boxplot corresponds to applying one of the four treatments of breakoff to the simulated breakoff datasets. Recall the four treatments tested in the study are: **IG** - Ignore breakoff, **NR** - Treat breakoff as nonresponse, **CW** - Combined weighting, and **MI** - Multiple imputation.

Given that there are 200 simulated breakoff datasets, each boxplot therefore consists of 200 survey estimates. The black solid line in the middle of the box represents the median survey estimate. The red dashed line is the benchmark value, which is calculated as the mean of the corresponding statistic of interest across 200 synthetic population data.

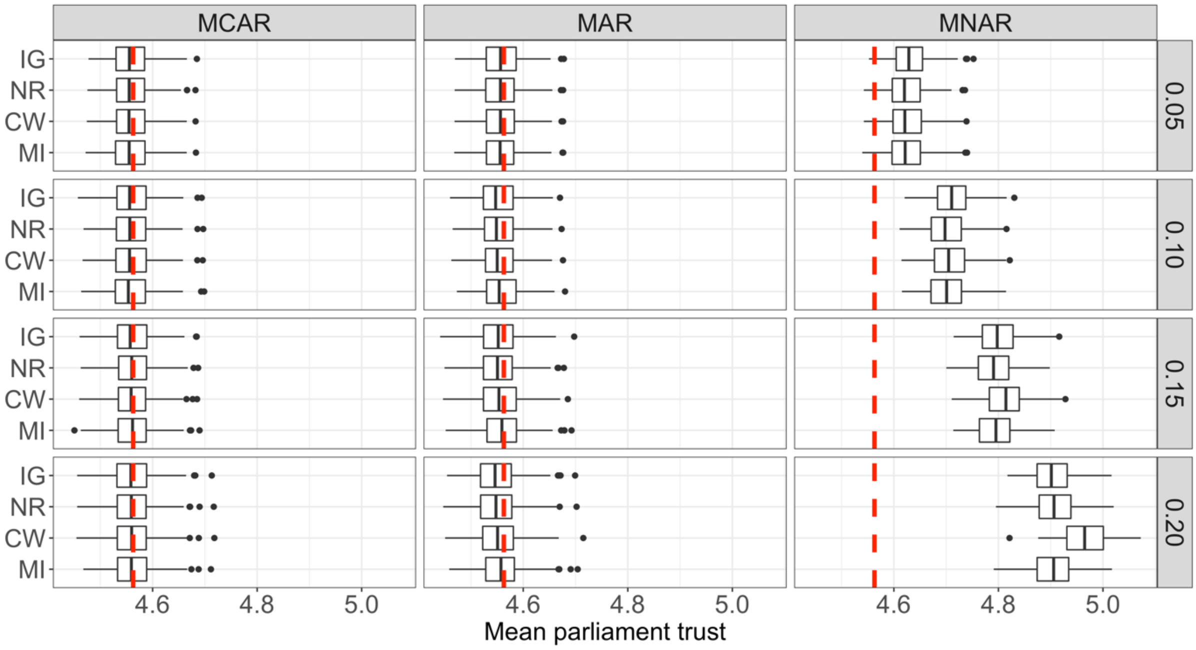


Figure C.1 Distribution of the mean parliament trust, estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

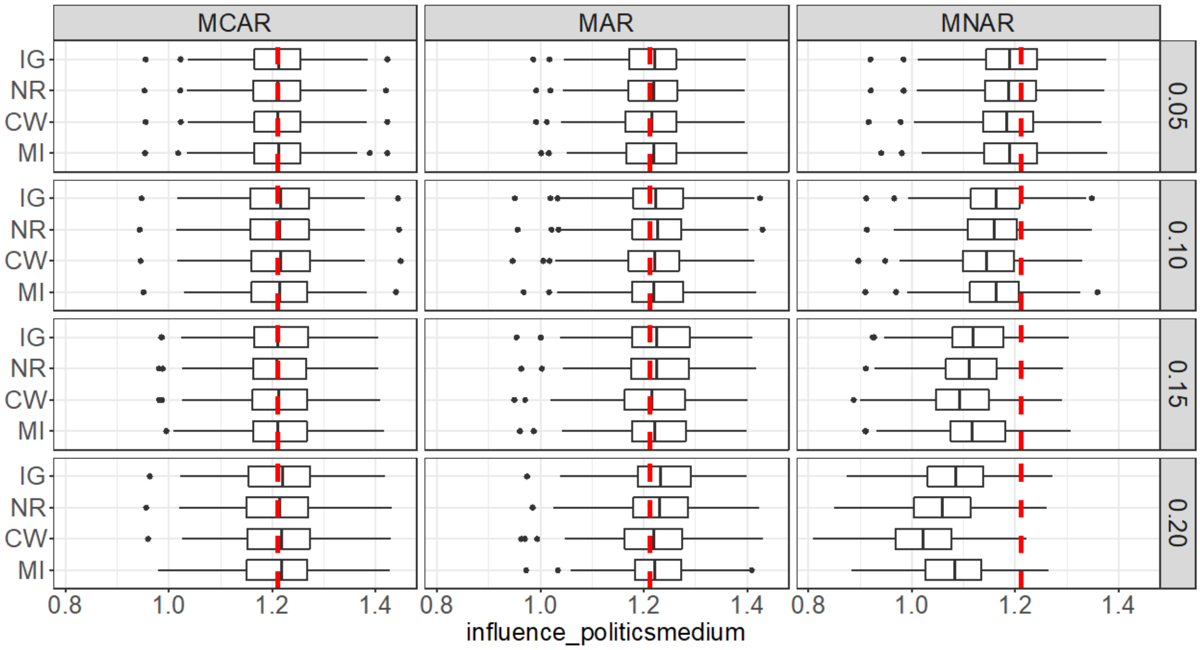


Figure C.2 Distribution of the model coefficient of perceived political influence (medium vs. low) on the parliament trust, estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

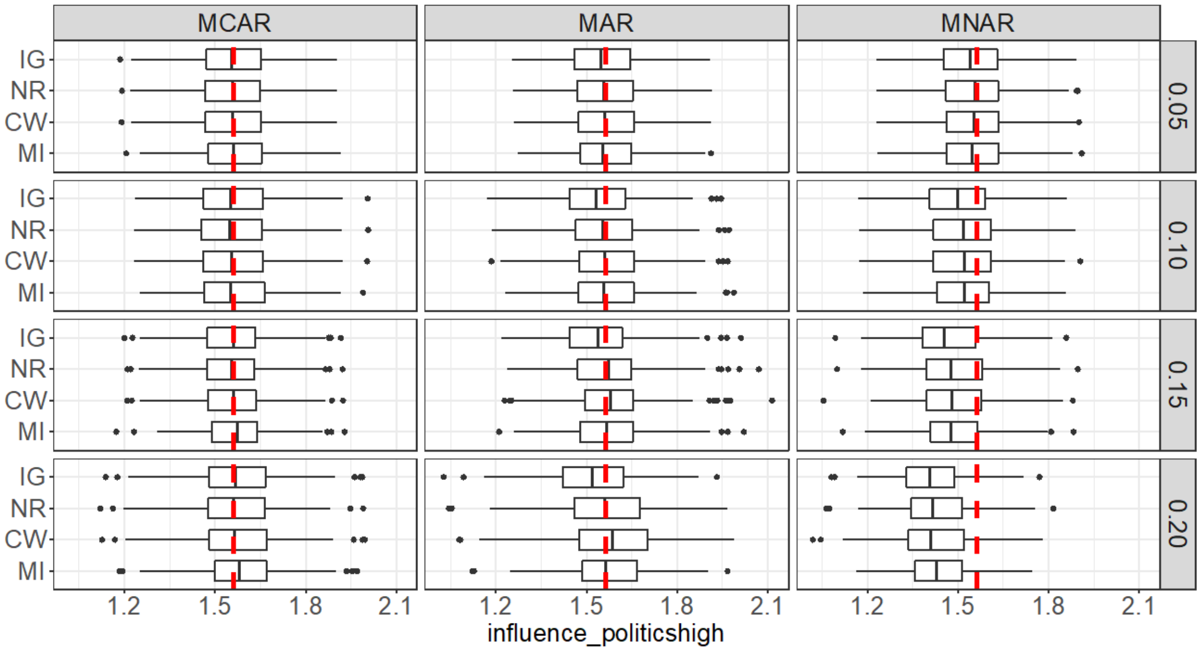


Figure C.3 Distribution of the model coefficient of perceived political influence (high vs. low) on the parliament trust, estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

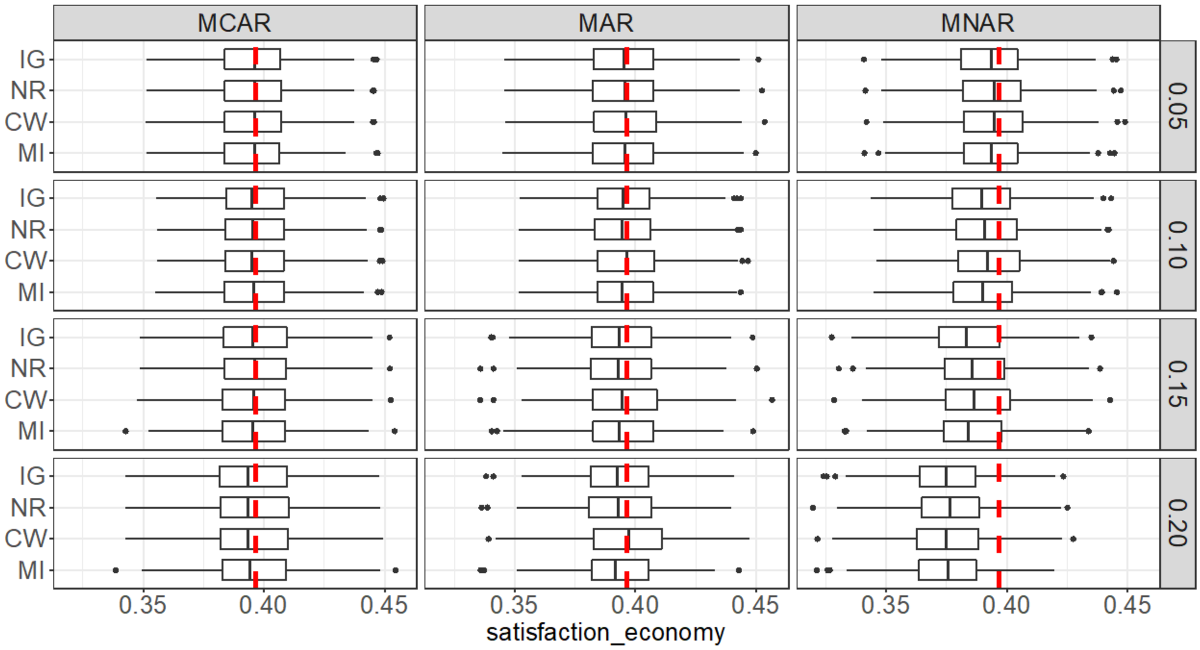


Figure C.4 Distribution of the model coefficient of satisfaction with economy performance on the parliament trust, estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

Appendix D

This appendix shows the result of Root Mean Squared Error. It is calculated using the equation below:

$$RMSE = \sqrt{\frac{\sum_{i}^{200}(x_i - X)^2}{200}}$$

The survey estimate is represented by x_i (i = 1, …, 200), which is derived by applying the four treatments of breakoff to each of the 200 simulated breakoff datasets. The benchmark value (denoted as *X*) is calculated by taking the average of the corresponding statistic of interest from the 200 synthetic population data.

Figures below display Root Mean Squared Error for the mean of parliament trust as well as the model coefficients regarding the impact of people's perceived political influence and satisfaction with country's economy on their parliament trust. The figures are separated by breakoff rates (right, row) and breakoff mechanisms (column). Each point in the figures represents the RMSE value resulted from using a specific method for dealing with breakoffs. The exact RMSE values are provided next to the corresponding points. Due to rounding, some points located differently on the horizontal scale might have the same RMSE values next to them.

Recall the acronyms in the plot correspond to the four breakoff compensation methods tested in the study. **IG**: Ignore breakoff, **NR**: Treat breakoff as nonresponse, **CW**: Combined weighting, and **MI**: Multiple imputation.

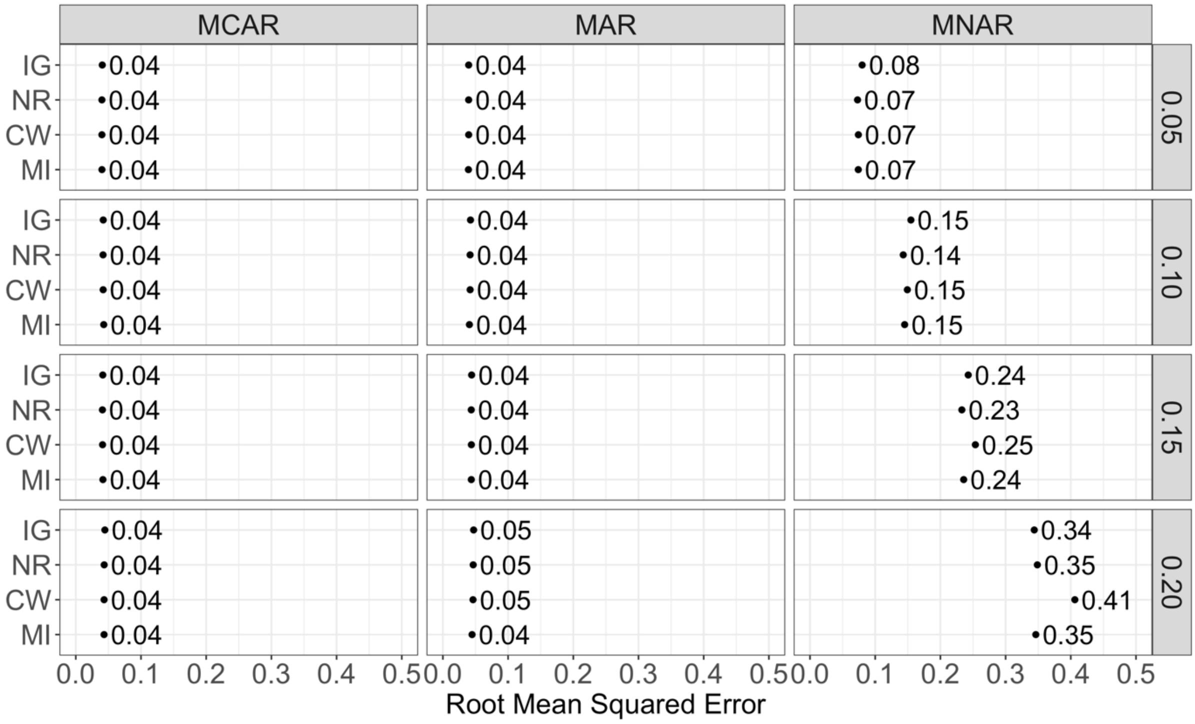


Figure D.1 Root Mean Squared Error of the mean parliament trust, estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

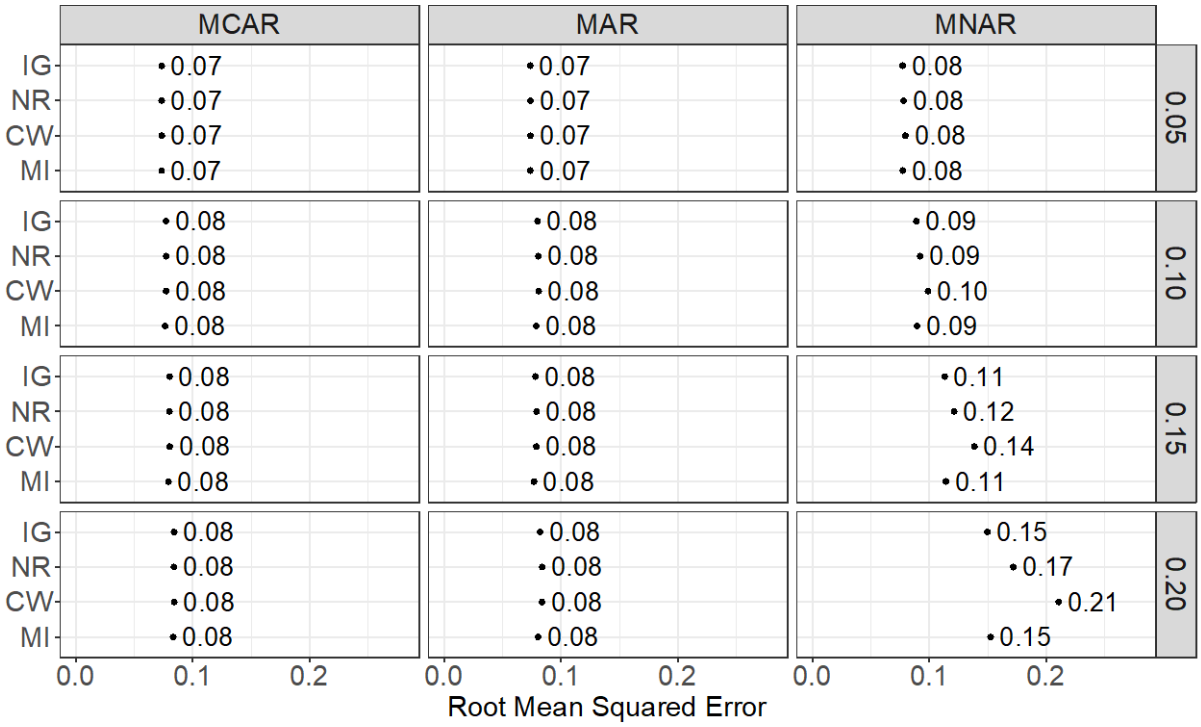


Figure D.2 Root Mean Squared Error of the model coefficient corresponding to the perceived political influence (medium vs. low), estimated using four breakoff compensation methods

(left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

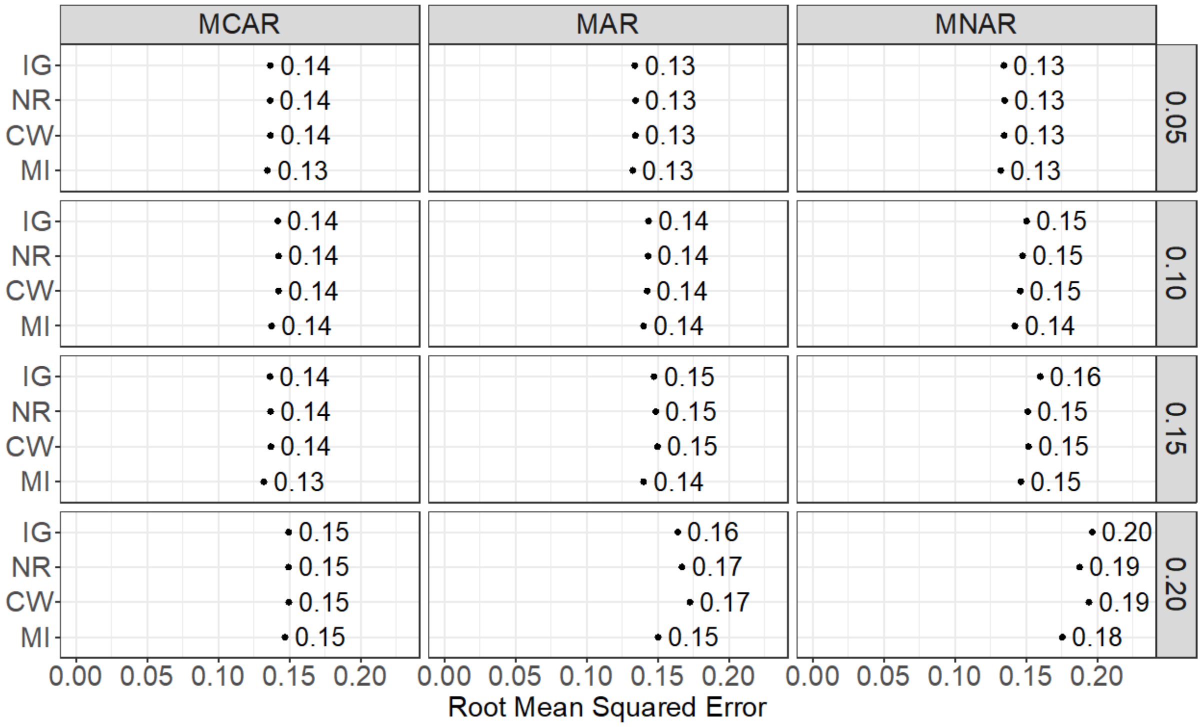


Figure D.3 Root Mean Squared Error of the model coefficient corresponding to the perceived political influence (high vs. low), estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).

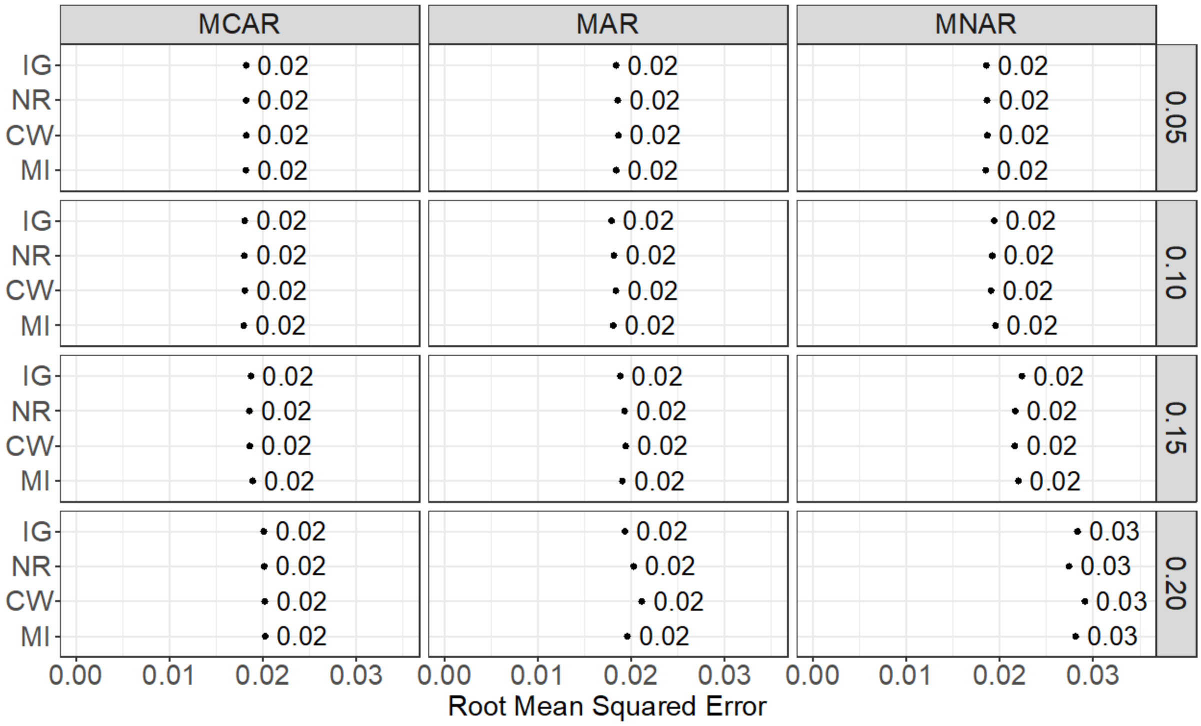


Figure D.4 Root Mean Squared Error of the model coefficient corresponding to the satisfaction towards country's economy performance, estimated using four breakoff compensation methods (left, rows) under three missing data mechanisms (columns) and four breakoff rates (right, rows).