Adaptive data collection using the paradata of the previous survey

**Keywords:** mixed-mode survey, logistic regression model, response profile

# Introduction

We present a method in which interviewer resources can be used more efficiently in mixed-mode survey. We focus on mixed mode surveys that include computer-assisted telephone interview (CATI) and web modes.  Our idea is to use the dataset of the previous survey and construct propensities to participate in the survey on the web (quickly or after the interviewer contacts) or on the CATI. The purpose of the propensities is that before a new survey begins, the propensities can be imputed to the sample. Using the propensities, potential web respondents are known, which allows interviewers to focus more effectively on those who are likely to prefer a telephone interview.

Many European statistical agencies have started to use mixed mode surveys, including the web option, for cost reasons [1]. Adaptive data collection is another way to reduce costs [2].  Most CATI interviews are conducted with a system that produces paradata containing information about calling times and call amounts per sample person. Web mode also produces information such as the questionnaire completion date. Adaptive data collection is possible to implement with these datasets. Combining other linkable administrative data sources and paradata, it is possible to model and predict response behavior and categorize respondents into different groups based on their response behavior. We use five categories for responsive behavior: easy (quick) web respondent, challenging web respondent, CATI preferer, refuser and non-contact. We construct propensities to belong to these groups.

In this paper, we study various questions on using the propensities as guidance for data collection. We examine e.g. does the propensities depend on the theme of the data collection, what is the relationship between different response behavior profiles and overall response propensity, and how to impute the propensities from one dataset to another. Ultimately, we try to find out whether using propensities produce into higher response rates or better representativeness among different groups. We use both mixed-mode pilot and production data from the Labour Force Survey (LFS) and mixed-mode pilot data from EU-SILC (shortly SILC) from 2019-2021.

# Methods

## Data and paradata

The SILC data comes from the mixed-mode pilot study conducted in the spring 2020. The data consists of 10 000 households and 3000 target persons and contains the first pilot round. The response rate was 37%. We also use administrative records linked to the original sample, which contains information on demographic variables such as age, gender, profession etc.

The LFS data comes from similar mixed mode pilot study from 2019 and it consists of the second and third rounds of the study. The sample size was 3200 and the response rate was 42%. Both pilot studies had a similar data collection design, where the respondents first received a letter inviting them to participate on the web. In LFS the letter could also be sent by e-mail or text if the respondent was contacted that way on the first round. After a certain period, the interviewers started to contact the non-respondents by calls, texts and e-mails and continued contacting until the end of data collection period. The interviewers contacted the non-respondents randomly, without priorizing calls.

The paradata contains information on each contact including contact time and its result. From the paradata we can study e.g. number of calls, messages or letters to each person, the result of every contact time and the final result (answering on the web, on the phone or reason for non-response).

## Construction of propensities

Using the paradata, we divide the LFS and SILC to following groups. Easy web respondents are persons who have completed the questionnaire on the net quickly, before any calls have been made. Challenging web respondents are persons who have completed the questionnaire on the net after one or more calls have been made. CATI respondents are persons who have participated in the survey on the phone. Refusers are persons who are non-respondents after calls. Non-contacted non-respondents are non-respondents who have not been contacted by the interviewer. The respondents belong to the first three groups, and the non-respondents belong to the fourth or fifth group. Each sample person belongs to only one group.

**Table 1. Groups and their proportions in the LFS and SILC.**

|  |  |  |
| --- | --- | --- |
| Group | LFS: Proportion of the sample (%) | SILC: Proportion of the sample (%) |
| Easy web respondents | 6.0 | 11.8 |
| Challenging web respondents | 16.9 | 13.7 |
| CATI respondents | 22.7 | 10.5 |
| Refusers | 36.9 | 24.3 |
| Non-contacted non-respondents | 17.5 | 39.6 |

We analyze belonging to these groups with a logistic regression model.  The idea is same as when response propensities are calculated from a response propensity model. In that model, the dependent variable is a binary response indicator (1 = respondent, 0 = non-respondent). We determine binary variables based on the five groups. For example, variable easyweb (1 = easy web respondent, 0 = other) determines whether a person is easy web respondent or not. We determine similar binary indicators for all five groups. Explanatory variables are register variables such as gender, age group and level of education, municipality group, language, the existence of an email address and occupation classification. The logistic regression models construct propensities to belong the easy web respondents, challenging web respondents etc.

Table 1 shows that when the datasets are divided into five groups, the proportions are quite low. For example, the easy web respondents of the LFS represent only 6.0% of the sample. Then it is more difficult to find explanatory variables to the logistic regression model.

## Imputation

The propensities can be imputed to a new sample using register variables. We use a donor-recipient method where donor and recipient are similar according to register variables. For this method, the both datasets must contain same register variables with similar classifications. The same method has also been used earlier when response propensities have been imputed to a new sample [3].

The imputation uses the same register variables that have been explanatory variables in logistic regression models. Note that the regression models may have different explanatory variables when it comes to easy web respondents, challenging web respondents etc. Thus, the five propensities are imputed separately. For example, suppose that the logistic regression model for easy web respondents contain three explanatory variables: gender (2 classes), education (3 classes) and language (3 classes). This means that the dataset contains 18 different easy web propensities. For example, easy web propensity for a Finnish speaking woman with a higher education degree is 0.114. People belonging to this group in the new sample will get imputed easy web propensity 0.114. This imputation method is sensitive to the number of imputation variables. If there are a lot of variables, imputation is difficult to implement. [3]. Therefore, the logistic regression model should contain about three best explanatory variables. Then the imputation should work well.

## Use of propensities in a new data collection

When the propensities have been imputed to the new sample, the propensities can be utilized in different ways in the data collection. People who have higher easy web propensity need fewer calls from the interviewers, because they are expected to participate in the survey independently on the web. The interviewers can try to recommend web mode to people who have higher challenging web propensity. If the interviewers succeed in this, they have less CATI interviews and they can better focus on the other persons in the sample. It is also useful to know potential CATI respondents, refusers or hard-to-contact people, because then the data collection can be planned in a new way.

# Results

We calculated propensities for the SILC, for the same groups as in Table 1. We used the same model for all dependent variables. The explanatory variables were gender, age group, education, municipality group and some interactions of these variables. We imputed the propensities to the LFS using the explanatory variables of the model in imputation. We also calculated propensities using the LFS dataset and same models and compared the propensities to imputed propensities by calculating correlation coefficients.

In the LFS and the SILC, gender and age distributions for the different respondent profiles were quite similar. This suggests that the response behavior is not linked to the theme of the data collection or to the length of the questionnaire. However, some differences were observed among profession and education classes.

In the both datasets, the explanatory variables had 71 combinations. This means that the models produced 71 different propensities for each response profile. The correlation between imputed propensities and LFS propensities varied. The correlation was 0.54 for easy web respondents, 0.74 for challenging web respondents, 0.92 for CATI respondents, 0.18 for refusers and 0.79 for non-contacted non-respondents. This means that the CATI respondents were quite similar according to auxiliary variables in the SILC and LFS, but the refusers were different. This suggests that the propensities should be jointly classified to create a trustworthy guiding parameter for the data collection.

We also constructed response propensities to the LFS and SILC to study connection between response profiles and response propensities. According to correlation coefficients, connection between response propensity and answering on the web is not clear. For example, a respondent with higher web propensity (easy web propensity or challenging web propensity), may have low response propensity, or vice versa.

# Conclusions

We have presented an adaptive data collection method which helps to use interviewer resources more effectively in a mixed-mode survey. We have constructed propensities for response behavior and presented how to impute the propensities to a new sample. When potential web respondents are known before a data collection, the interviewers can target their calling to potential CATI respondents, refusers and non-contacted non-respondents. Using the propensities also provides opportunities when, for example, more experienced interviewers can take potential challenging respondents.

It seems that the theme of the survey or the length of the questionnaire does not affect theresponse behavior. The SILC data is from the first round, and the LFS data is from the second and third round. The response profiles also seem to remain quite similar between panel rounds.

In our study, different response profiles did not correlate strongly with response propensities. For this reason, care is required when using propensities presented in this paper. Response propensities should be considered at the same time. The interviewers need to call all sample persons. They cannot suppose, for example, that easy web respondents participate in the survey without calls. An interesting question is if our method produces a better response rate or reduces nonresponse bias. When interviewer resources are utilized more efficiently, it should improve the response rate. The effect on the nonresponse bias is more difficult to predict. This method should be tested in practice. We aim to use our findings in the LFS mixed-mode data collection in 2021 and present some preliminary results in the NTTS2021 conference.

# References

1. E. de Leeuw, Z. Suzer-Gurtekin and J. Hox, How to Design and Implement Mixed Mode Surveys in Cross National Surveys: Overview and Guideline. Second International Conference on Survey Methods in Multinational, Multiregional and Multicultural Contexts (3MC) (2016), Chicago.
2. K. van Berkel, Adaptive data collection at Statistics Netherlands with an application to the Health Survey. Conference of European Statisticians (2018), Geneva.
3. M. Honkala, Estimation of Response Propensities Using the Previous Survey,      International Journal of Statistical Distributions and Applications, 5 (2019), 5-9.